IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 December 2022, accepted 11 January 2023, date of publication 16 January 2023, date of current version 24 January 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3237025*

SURVEY

# Domain Adaptation: Challenges, Methods, Datasets, and Applications

**PEEYUSH SINGHAL[1], RAHEE WALAMBE[1,2], (Senior Member, IEEE), SHEELA RAMANNA[3], AND KETAN KOTECHA[1,2]**

[1]Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, Maharashtra 412115, India
[2]Symbiosis Centre for Applied Artificial Intelligence (SCAAI), Symbiosis International (Deemed University), Pune, Maharashtra 412115, India
[3]Department of Applied Computer Science, The University of Winnipeg, Winnipeg, MB R3B 2EP, Canada

Corresponding author: Ketan Kotecha (director@sitpune.edu.in)

Sheela Ramanna's research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants program (#194376).

**ABSTRACT** Deep Neural Networks (DNNs) trained on one dataset (source domain) do not perform well on another set of data (target domain), which is different but has similar properties as the source domain. Domain Adaptation (DA) strives to alleviate this problem and has great potential in its application in practical settings, real-world scenarios, industrial applications and many data domains. Various DA methods aimed at individual data domains have been reported in the last few years; however, there is no comprehensive survey that encompasses all these data domains, focuses on the datasets available, the methods relevant to each domain, and importantly the applications and challenges. To that end, this survey paper discusses how DA can help DNNs work efficiently in these settings by reviewing DA methods and techniques. We have considered five data domains: computer vision, natural language processing, speech, time-series, and multi-modal data. We present a comprehensive taxonomy, including the methods, datasets, challenges, and applications corresponding to each domain. Our goal is to discuss industrial use cases and DA implementation for those. Our final aim is to provide future research directions based on evolving methods and results, the datasets used, and industrial applications.

**INDEX TERMS** Artificial intelligence, computer vision, deep neural network, domain adaptation, multi-modal data, natural language processing.

## I. INTRODUCTION

Leon C. Megginson summed up Charles Darwin's work [1] by saying, ''It is not the strongest of the species that survives, not the most intelligent that survives. It is the one that is most adaptable to change''. The same thing can also be said about technology. The workhorse of Machine Learning (ML) and Artificial Intelligence (AI) – supervised learning has a severe disadvantage in that it works well when samples for training and testing both belong to the same distribution and are independent and identically distributed (i.i.d.). Domain Adaptation (DA) is a special case of Transfer Learning (TL), which supports and solves real-world (including in the wild) challenges by effectively applying the model trained on one dataset (source) for testing on another domain (target) with different distribution.

Domain Adaptation (DA) is increasingly acquiring traction from academia and industry since it promises the practical and evolving side of AI and ML. DA, in many ways, mimics how humans learn and adapt to the real world around them. In practice, we see that the supervised learning model's accuracy (or another performance metric) is not transferrable for the same tasks to datasets not used as part of the training. The primary reason for this failure is a deviation from an assumption- the source and target domain data are drawn from the same distribution. The problem is further accentuated when we understand that acquiring labeled data is time-consuming, costly, and at times, infeasible – which means

The associate editor coordinating the review of this manuscript and approving it for publication was Michele Nappi.

the state-of-the-art models are limited to only some academic datasets. The performance degradation is caused by domain shift (domain gap or dataset bias): the difference in data distributions between source and target domains. DA is a field of AI that aims to alleviate as far as possible the impact of domain shift and ensures that the models perform well in the target domain after being trained on the source domain. The target and source domains should have some similarities (e.g., features) for a meaningful adaptation.

DA provides an attractive option for Deep Learning (DL) – which, more often than not, provide high performance over shallow learning or classical learning algorithms. DA negates the vast amount of labeled data requirements in the target domain and typically uses available (labeled) data in the source domain, a boon to data-hungry supervised DL algorithms. Realistically, there is an excessive amount of unlabeled data available, but labeled data is scarce. Some techniques have been tried to better the performance metric of deep networks by using more data (labeled) from the target domain, including better/alternative architectures and backbones, use of normalization layers (e.g., Instance Normalization (IN) [2], Batch Normalization (BN) [3]), data generation and data augmentation, etc. By far, DA appears to provide a more robust alternative to all the mentioned techniques.

Initial work on DA is related to shallow (or classical) learning. With DL more prevalent in recent years, the focus of research shifted to DA in DL. The invention of GAN [4], Attention and Attention-based Transformers [5] have boosted various DA in DL methods. The research direction and focus now is to solve real-world and practical setting problems with the latest methods and techniques (e.g., few or zero-shot, self-supervised learning, meta-learning, etc.) and with real-world data situations (e.g., multi-modal data, multi-domain, continuous/ incremental domains, and data restriction, etc.). This survey does not focus at length on Domain Generalization (DG), a related area where information about the target domain is unknown.

A number of survey papers on DA are reported. The primary difference between this and the previous works is threefold; this survey encompasses various data domains instead of only focusing on a specific (text/image-based) modality. Secondly, the survey is conducted with a primary focus on the applications of DA in these data domains – the challenges faced and how those can be mitigated using DA. Thirdly, it tries to understand the application of DA approach across data domains/modalities and also tries to understand what makes a particular DA approach data domain specific. In summary, the primary goals of this work are:

1. To provide a joint perspective and recent updates of domain adaptation in five deep learning data domains – Visual or Computer Vision (CV), Natural Language Processing (NLP), speech, time-series, and multi-modal domains. Most of the previous surveys only focused on the visual domain (CV) or NLP domain and missed out on areas of cross-pollination. This survey, we believe, for the first time, discusses DA in multi-modal data settings. To understand data domain (CV, NLP, speech, time-series, multi-modal) specific DA methods and techniques and ones that are used across data domains.
2. To compile a list of existing and emerging DA datasets and tasks in five data domains.
3. To review recent DA methods and techniques for more practical DA settings like learning with fewer data, learning on the go, continuous adaptation, presence of domain or category gap, etc., across data domains.
4. To understand challenges and issues that hinder the adoption of DA. Based on these challenges and issues, research directions are also provided. These challenges and issues also provide research direction.
5. Understanding and reviewing industrial use-cases where DA has been employed and appreciating use-cases where DA if deployed, would provide rich dividends.

*Organization* ***of paper:*** Pictorial view of the organization of the paper can be seen in Figure 1. For completeness, the survey also briefly discusses the background, definition, and theory of DA in section II and then discusses DA in shallow or classical learning in section III. DA in DL is discussed in section IV; this section also focuses on more practical DA settings. Datasets used in five data domains and observations are mentioned in section V. Challenges and issues being worked on in this field are mentioned in section VI. Section VII looks at common and specific DA use-cases across industries and provides a perspective on how DA can be helpful. Section VIII provides the future research frontiers. The paper is concluded in section IX.

## II. BACKGROUND

This section aims to succinctly provide the formal definition of DA, the categories of transfer learning and domain adaptation, and a theoretical foundation of domain adaptation.

### A. FORMAL DEFINITION OF DOMAIN ADAPTATION

Let there be a source domain $D^s$, composed of a feature space $\chi^s$ and marginal probability distribution $P(X^s)$ such that $D^s = \{\chi^s, P(X^s)\}$. Also, there exists a sample set $X^s = \{x_1^s, x_2^s, \ldots, x_n^s\}$ and corresponding labels $Y^s = \{y_1^s, y_2^s, \ldots, y_n^s\}$ from $\Upsilon$.

Similarly, there is a target domain $D^t$, composed of a feature space $\chi^t$ and data with marginal probability distribution $P(X^t)$ such that $D^t = \{\chi^t, P(X^t)\}$. Also, there exists a sample set $X^t = \{x_1^t, x_2^t, \ldots, x_n^t\}$ and corresponding labels $Y^t = \{y_1^t, y_2^t, \ldots, y_n^t\}$ from $\Upsilon$.

Sometimes, labels in the target domain are unavailable (case of unsupervised DA) or only a few are available (case of semi-supervised DA), or no data at all is available in the

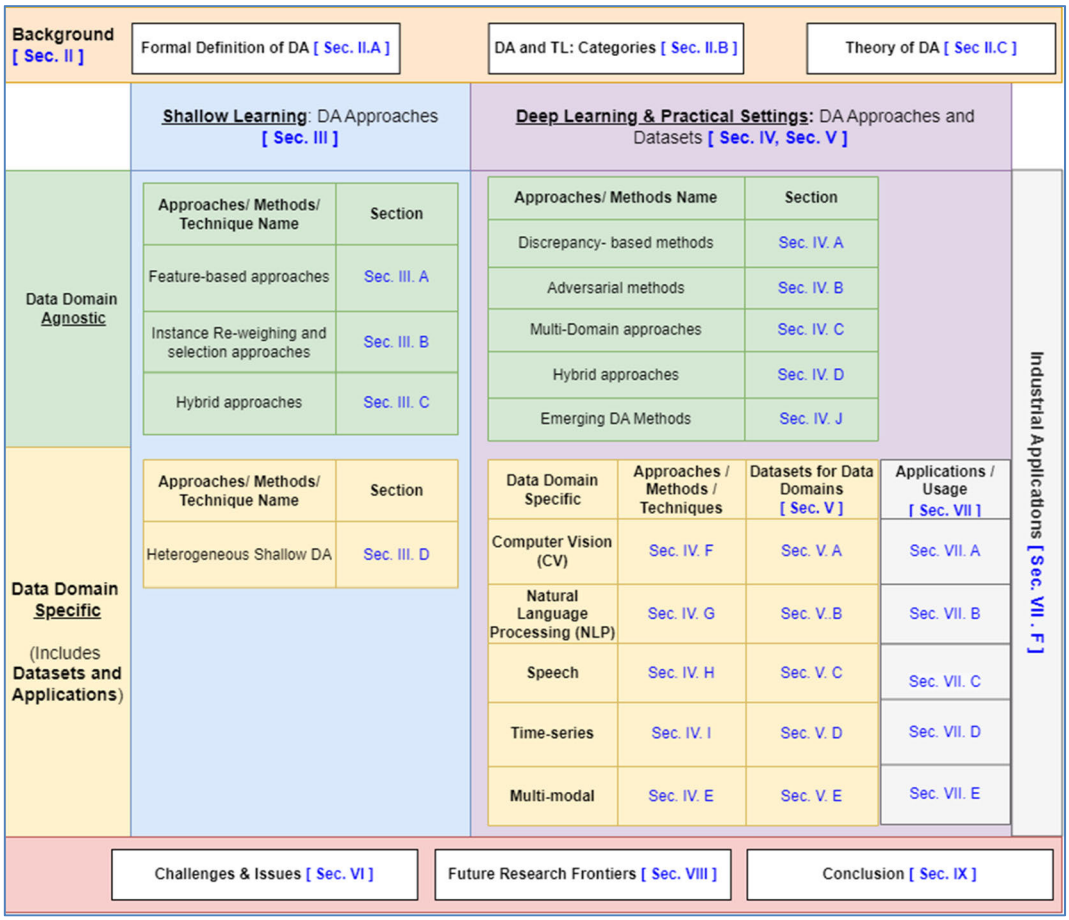


**FIGURE 1. Organization of the paper. Paper flow is from top to bottom and right to left. Industrial applications (Sec. VII-F) include applications of both shallow and deep learning domain adaptations. Data Domain (CV, NLP, Speech, Time-series, Multi-modal) specific approaches / methods, Datasets and Applications are contained in subsections of Sec. IV, Sec. V, Sec. VII respectively.**

target domain (case of domain generalization or zero-shot DA). Supervised or unsupervised DA refers to labels in the target domain being available or not for training. There exists a domain shift between $D^s$ and $D^t$. The task in the source domain is: $\mathcal{T}^s = \{\Upsilon^s, P(Y^s|X^s)\}$ and the target domain is: $\mathcal{T}^t = \{\Upsilon^t, P(Y^t|X^t)\}$.

In the case of DA, if there exists a mathematical model $f : X^s \rightarrow Y^s$. If, $\mathcal{T}^s$ is related to $\mathcal{T}^t$, and the same model $f$ also works for $X^t \rightarrow Y^t$ with a minimal error or acceptable error, the model $f$ has adapted to the target domain $D^t$ and source domain $D^s$.

### B. CATEGORIES OF TRANSFER LEARNING AND DOMAIN ADAPTATION

The seminal work on DA by Pan and Yang [6] mentions that DA is a specific case of transfer learning (TL). The commonality between DA and TL is that some learning based on source domain data is utilized for the task in another. Hence it is beneficial to understand different instances/types of TL.

A. Based on the feature set and data distributions, there are two types of transfer learning approaches (refer to Table 1)
B. Based on the task difference and the corresponding source and target domain data (refer to Table 2)

Figure 2 shows DA categories based on various source and target domain characteristics. DA work typically falls into homogeneous and transductive TL. However, in the recent past, there have been reasonable attempts to focus on heterogenous DA. DA can be categorized based on the availability of labels in the target domain (refer to Table 3 and Table 4).

DA can also be categorized based on the label (classes) in domain and source data (refer to Table 4).

Typically, domain classification represents the scenario when there is only a single source domain. The adaptation is to another single-target domain (called single-target DA). However, recently, DA to multiple target domains (called multi-target DA) is also reported. Adaptation from multiple source domains (called multi-source DA) has also been researched.

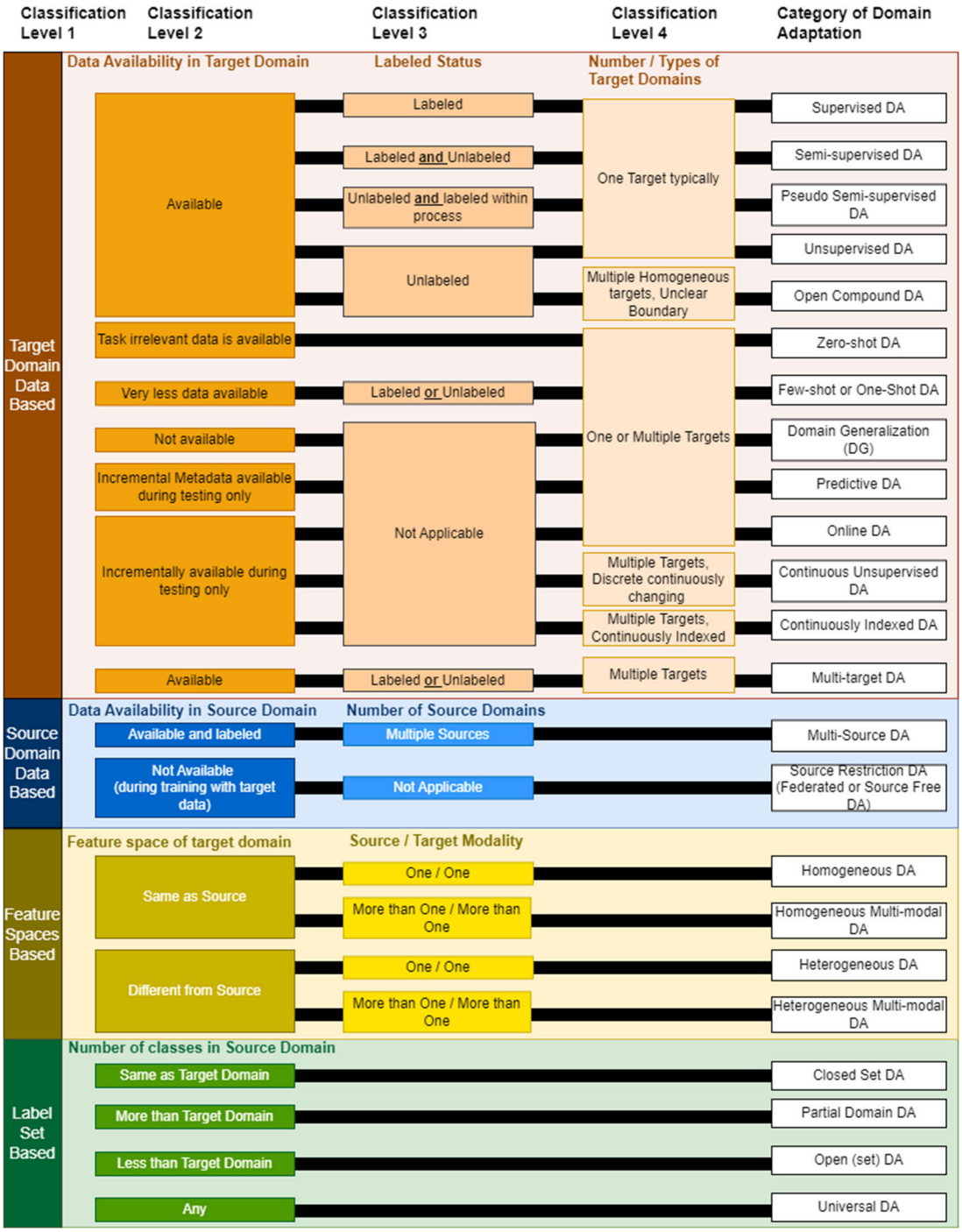


**FIGURE 2.** Overview of DA categories based on data characteristics (availability, feature space, number of classes) of the source and target domains. Subclassification within each class leads us to specific category of DA.

Until more recently, the DA focus was on reducing the dependency of labeled instances of data in the target domain; now, researchers are also focusing on reducing the dependency of data itself in the target domain. Few-shot DA, single-shot DA, and zero-shot DA are examples of efforts to incrementally reduce the requirement of target domain data. Predictive DA uses metadata in the target domain to adapt. Domain generalization (DG) can be seen like zero-shot DA, but it is bereaved of knowing anything about the target; however, more robust DG methods should also include some essence of multi-target DA, Universal DA. DA techniques also focus on the absence of source data during the DA

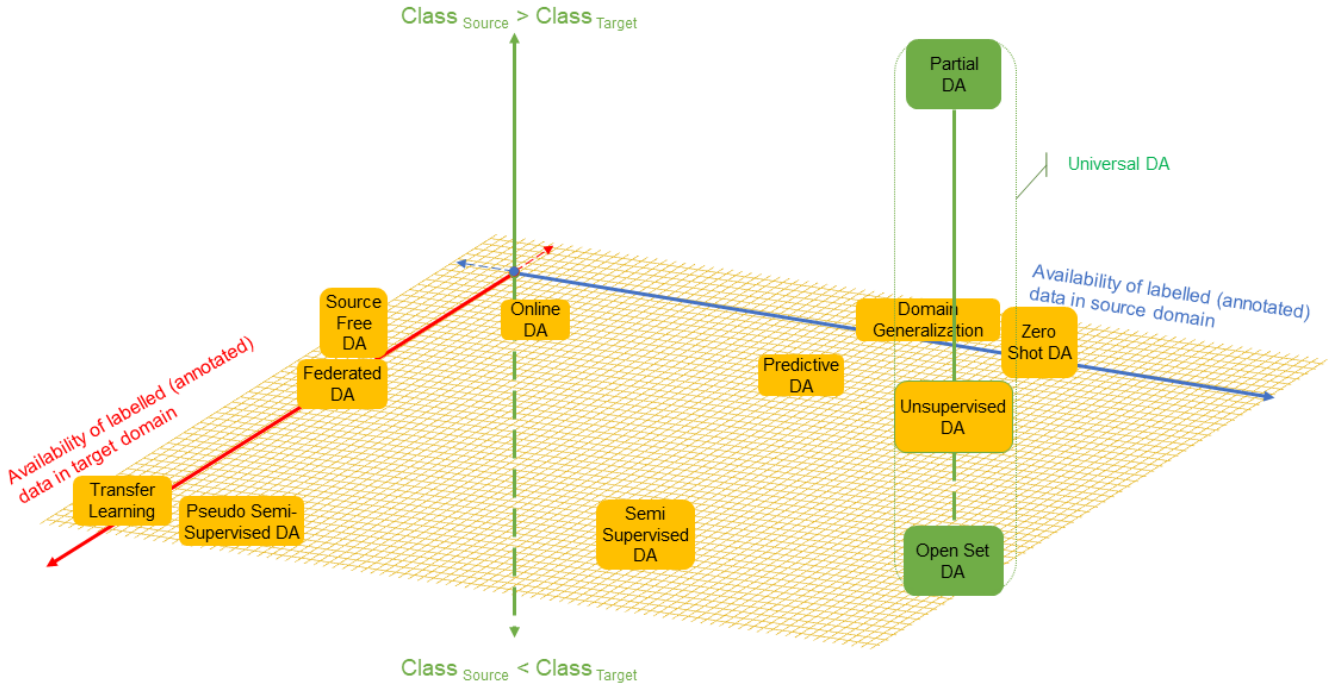


**FIGURE 3.** Domain adaptation categories plotted based on the availability of annotated data in the source and the target domain (forming aa horizontal plane) and category (class) set difference in source and target (forming the vertical axis), enhanced and adapted from Tommasi [7].

**TABLE 1.** Transfer learning (/domain adaptation) categories based on data.

| S. No. | Case | Type |
|---|---|---|
| 1 | In this case, the domain shift is based on two domains having the same input feature space ($\chi^s = \chi^t$), but the shift is because of different data distributions, i.e., $P(X^s) \neq P(X^t)$. Domain dimensions are also the same. | Homogeneous Transfer Learning / Domain Adaptation |
| 2 | In this case, the domain shift is due to the difference in the input feature space ($\chi^s \neq \chi^t$). It may be possible that the modalities may also differ – e.g., one is text, and one is an image. | Heterogeneous Transfer Learning / Domain Adaptation |

**TABLE 2.** Transfer learning categories based on task differences and data.

| S. No. | Case | Type |
|---|---|---|
| 1 | $D^s = D^t$ and $\mathcal{T}^s = \mathcal{T}^t$ | Typical Supervised Learning the training set is $D^s$, and the test set is $D^t$ |
| 2 | $D^s \neq D^t$ and $\mathcal{T}^s = \mathcal{T}^t$ | Transductive Transfer Learning **Domain Adaptation** (Further restriction is that labels space should be shared $Y^s = Y^t = Y$) |
| 3 | [$D^s = D^t$ or $D^s \neq D^t$] and $\mathcal{T}^s \cong \mathcal{T}^t$ | Inductive Transfer Learning Requires small data from the target to induce the model. |
| 4 | $D^s = D^t$ and $\mathcal{T}^s \neq \mathcal{T}^t$ | Generative Learning |
| 5 | $D^s \neq D^t$ and $\mathcal{T}^s \neq \mathcal{T}^t$, but $D^s$ is related to $D^t$ or $\mathcal{T}^s$ is related to $\mathcal{T}^t$ | Typically, unsupervised learning or Custom |

process – this may be due to privacy reasons (Federated DA) or plain unavailability (Universal source-free DA).

Tommasi [7] (refer to Figure 3) categorized different DA approaches based on the amount of data available and the number of classes in the source and target domain.

### C. THEORY OF DOMAIN ADAPTATION

The works of Ben-David and collaborators ([8] and [9]) looked at formulating the theoretical assumptions of the DA problem. They and future researchers were interested in finding out how real-world challenges deviate from theoretical assumptions. Ben-David et al. [9] calculated a bound on the DA error (empirical target error) for a semi-supervised case as (1).

$$\epsilon_T(\hat{h}) \leq \epsilon_T(\hat{h}_T^*) + (4\sqrt{(\alpha^2/\beta) + (1-\alpha)^2/(1-\beta)}) \times \left(\sqrt{\frac{\left(2d\log(2(m+1)) + 2\log\left(\frac{8}{\delta}\right)\right)}{m}}\right) + 2(1-\alpha)\left(\frac{\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T)}{2} + 4\sqrt{\frac{(2d\log(2(m'+1)) + 2\log(\frac{8}{\delta}))}{m'}} + \lambda\right) \quad (1)$$

**TABLE 3. DA categories based on the availability of labeled target domain data.**

| S. No. | Case | Type |
|---|---|---|
| 1 | In this case, labels for both domains are available, i.e., $Y^s = \{y_1^s, y_2^s, \ldots, y_n^s\}$ and $Y^t = \{y_1^t, y_2^t, \ldots, y_n^t\}$ are both available | (fully) Supervised DA |
| 2 | In this case, labels for both domains are available; however, only a minimal quantity of target domain labels is available, i.e., $Y^s = \{y_1^s, y_2^s, \ldots, y_N^s\}$ and $Y^t = \{y_1^t, y_2^t, \ldots, y_n^t\}$ are both available and $N \gg n$. | Semi-Supervised DA |
| 3 | In this case, labels for only source domains are available, i.e., $Y^s = \{y_1^s, y_2^s, \ldots, y_n^s\}$ and $Y^t = \{\emptyset\}$ | Unsupervised DA |
| 4 | As in the case of Unsupervised DA, only labels for the source domain are available at the start of the DA process. However, labels (accurate or pseudo) are obtained during the start of the "core" DA process, and the DA process is, therefore, a Pseudo-Semi-Supervised DA<br>• Active DA applies different sample selection methods with the intention of selecting diverse yet complex samples on target data. The labels for the selected target data are then obtained. The labels obtained are accurate.<br>• Pseudo-label DA does not focus on diversity and complexity in sample selection techniques (if at all followed). One of the techniques to do pseudo-labeling of target data is by applying the source model to the subset of target data to obtain labels (another may be self-supervised learning). The labels obtained are not accurate. | Pseudo-Semi-Supervised DA |

**TABLE 4. DA categories based on label set in target domain data. $C_s$&$C_t$ Represent label set (not the number) in sources & target domain, respectively.**

| S. No. | Case | Type |
|---|---|---|
| 1 | $C_s = C_t$, Same labels (classes) in both domains. | Closed Set DA |
| 2 | $(C_s \cap C_t) \subset C_s$, the common labels (classes) are a subset of the source domain. There are no out-of-set classes present. | Partial Set DA |
| 3 | $(C_s \cap C_t) \subset C_t$, the common labels (classes) are a subset of the target domain. There are out-of-set classes (not seen by the model before) present. | Open Set DA |
| 4 | There is no prior knowledge of the labels (classes) in the target set. | Universal DA |

In (1), $\epsilon_T\left(\hat{h}\right)$ is the empirical target error, $\alpha$ is a linear combination of errors in sources and target domains, $m$ is sample size with $(1-\beta)m$ points are drawn from source domain while $\beta m$ drawn from target. $\delta$ signifies the probability. $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T)$ is $\mathcal{H}\Delta\mathcal{H}$ divergence (or simply $\mathcal{H}$-divergence) between source and target samples, and $\lambda$ is predictor error.

Researchers involved in finding the basis of the theoretical formulation of DA mention three primary conditions required for DA.

1) ***Covariate Shift*:** Conditional label distribution is the same - $P(Y^s|X^{\mathbf{s}}) = Q(Y^t|X^{\mathbf{t}})$, between source and target distributions.
2) ***Somewhat similar distributions*:** Source and target distributions should be somewhat similar, i.e., $P_x$ *and* $Q_x$ must be similar. Typically (as in [9]), $\mathcal{H}$-Divergence is used to understand the difference in distribution.
3) ***Joint error minimization*:** DA works to minimize the joint error on source and target.

However, various works - [8], [9], and [10] - then focused on unraveling the above conditions, which are not sufficient to guarantee a good DA in the real world. Zhao et al., in the theoretical study [10], concentrated on domain-invariant learning methods and proposed the removal of the joint error minimization condition mentioned before. Another theoretical basis for DA in DL was offered by Le et al. [11], explaining why it is possible to close the gap between domains in joint space.

## III. DOMAIN ADAPTATION IN SHALLOW (OR CLASSICAL) LEARNING

To grasp DA in DL, it is important to learn DA in shallow learning (or classical learning) to provide the chronology. Any work associated with DA that does not include DL is considered shallow learning and caters to the DA work that happened before the use of DNNs became more prevalent. The DA methodologies in both shallow and deep learning aim to strengthen the model somehow by using the features invariant to domains (also called domain-invariant) or transforming the target data into a form/space in which the model is trained to reduce the task error. However, given that most of the features of shallow learning are handcrafted and explainable, the features can be inspected separately. DA in shallow learning is mostly based on features (matching or alignment or transformation or augmentation) and less on data-instance based. Work done by Csurka [12] provided a comprehensive survey of shallow DA methods in the visual domain. This section extends Gabriela Csurka's work [12] by including frequently used shallow DA strategies in NLP, time-series and other data domains along with CV domain DA methods.

### A. FEATURE BASED APPROACHES

Feature-based approaches (refer to Table 5) are prevalent in both shallow (origin) and deep DA. The main idea associated with feature-based approaches (matching/ alignment/ transformation/ augmentation) is to find a shared feature embedding/representation by reducing the data distribution difference. An effort is taken in the approaches to preserve input data properties.

From Table 5, we observe two important aspects:

1) Maximum Mean Discrepancy (MMD) [14] is used by multiple methods (Table 5 column – "The criterion (/criteria) for distribution difference / Discriminative Methods") to understand the distance between source and target distributions.

Definition of MMD: If sample set $X^s = \{x_1^s, x_2^s, \ldots, x_n^s\}$ and $X^t = \{x_1^t, x_2^t, \ldots, x_m^t\}$ are from distributions $P(X^s)$ and $P(X^t)$ respectively, then MMD is defined in (2):

$$\widehat{MMD}\left(P(X^s), P(X^t)\right) = \left\| \frac{1}{n}\sum_{i=1}^{n} \varphi\left(x_i^s\right) - \frac{1}{m}\sum_{i=1}^{m} \varphi(x_i^t) \right\|_H^2 \tag{2}$$

where $\mathcal{H}$ is a universal RKHS, and $\varphi : \chi \to \mathcal{H}$

We see from the definition that:

a. MMD is non-parametric, which leads to a closed-form solution - a trivial solution.
b. MMD is dependent only on features and independent of classes and class labels and therefore supports unsupervised DA. In the case of semi-supervised or supervised, or pseudo-semi-supervised settings, class-conditioned MMD can be used to further improve DA.

Further, the use of the Kernel trick

$$k(x_i, y_j) = \varphi\left(x_i\right)^T \varphi\left(y_i\right) = \langle \varphi\left(x_i\right) \mid \varphi\left(y_i\right) \rangle_H \tag{3}$$

i.e., dependency on the inner product only simplifies the MMD estimation, as any distance between samples is the inner product, and an inner product can be represented as a kernel.

2. Use of Reproducing kernel Hilbert Space (RKHS):
   a. When data is transformed into sparse spaces (like RKHS), the chances that it is linearly separable are high.
   b. Representer theorem applied to the inner product would mean that the inner product of samples and the inner product of samples in RKHS are the same.

Therefore, transforming the samples to RKHS gives a distribution difference not only in RKHS but also in the feature space dimension.

### B. INSTANCE RE-WEIGHTING AND SELECTION APPROACHES

Another widely used strategy is instance re-weighing; the focus here is on the input data altogether and not on features. Further, the distribution difference is minimized by reweighting the source data for the task. The instance re-weighting approach is also called instance selection, as it leads to soft/hard selection of data.

Table 6 mentions the instance re-weighting and selection approaches. However, the re-weighting strategy does not help much when there is little overlap between the source and target domain. Little overlap leads to a small set of source domain examples assigned high weights – leading to a sub-optimal classifier as it tends to see a smaller number of samples effectively. However, in specific scenarios, as mentioned by Jong [23], re-weighting can provide a decision boundary closer to the optimal decision boundary of the target data.

### C. HYBRID APPROACHES

Hybrid approaches typically use both feature-based and instance re-weighting methods. An example of this is Transfer Joint Matching [24], wherein the feature matching is done by minimizing MMD in an infinite-dimensional reproducing kernel Hilbert space (RKHS). They also do reweight by minimizing the $l_2$-norm.

### D. DOMAIN ADAPTATION FOR HETEROGENOUS DATA (HETEROGENOUS SHALLOW DA)

As we have seen in previous subsections, the two DA domains were homogeneous, i.e., $\chi^s = \chi^t$; however, DA techniques have been applied to heterogeneous data (including multimodal data). In the case of heterogeneous data, we see $\chi^s \neq \chi^t$. Primarily, two transformation strategies are seen in Heterogenous Shallow DA – symmetric and asymmetric transformation.

When an attempt is made to project the source domain and target domain features to a common subspace (domain-invariant common latent subspace), the attempt to learn feature transformation is known as symmetric transformation. In asymmetric transformation, either source features or target features are transformed and aligned to target features or source features, respectively. An example of symmetric transformation is Heterogenous Feature Augmentation (HFA) [29]. At first, HFA transforms data (using projection matrices) from both domains into a common subspace. Then HFA augments transformed data, using two feature mapping functions, with the original features and zeros. SVM with hinge loss is applied to augmented features to learn the project matrices. Asymmetric Regularized Cross-domain Transformation (ARC-t) [30] is an example of asymmetric transformation. It uses a Gaussian radial basis function (RBF) kernel to learn asymmetric and non-linear transformation while mapping target data to source data. In ARC-t [30], it is mentioned that the power of ARC-t is that it can be applied to categories that were unavailable during training too.

Another perspective, according to Csurka [12], is that multi-view learning can be strongly related to heterogenous DA, in that multi-view solves the task by looking at features

**TABLE 5.** Feature based shallow DA approaches.

| Sr. No. | Approach | Data Domain | Example | Key Idea | Space Dimension | The criterion (/criteria) for distribution difference / Discriminative Methods |
|---|---|---|---|---|---|---|
| 1 | **Feature Matching**: The aim here is to find domain-invariant feature learning or representation. i.e., learning nothing specific to the domain yet learning specifically for the task. The reasoning behind this thought is that classifier trained on the source domain's domain-invariant features should work well in the target domain. | CV | Maximum Mean Discrepancy Embedding [13] | Common latent representation and properties of input data are preserved. The use of distance metrics is seen. | Lesser dimension than input data – Dimensionality reduction using PCA | Maximum Mean Discrepancy (MMD) [14] |
| | | CV | Transfer Component Analysis (TCA) [15] | Learning common transfer components underlying both domains. Focus on discovering components that do not cause a change in distribution across domains. The structure of data is preserved. | Higher than input data but not very high/infinite – RKHS | Maximum Mean Discrepancy (MMD) [14] |
| | | NLP | Structural Correspondence Learning (SCL) [16] | Prediction of "pivot features", features that are common to both the domains (with unlabeled data) – it uses auxiliary functions to construct the shared space and induce correspondence among different features. | low-dimensional real-valued feature space | Huber loss [17] |
| 2 | **Feature Augmentation:** Use of intermediate feature representations, which are cross-domain between source and target domain. These representations are typically label agnostic and therefore support both unsupervised, semi-supervised, supervised DA | NLP | Both source and target feature augmentation [18] | Augmentation of feature space of both source and target data, use that as input to any ML algorithm (e.g., SVM) | Augmented space is three times the feature space of the source/ target | Singular Value Decomposition (SVD) |
| | | CV | Geodesic Flow Sampling (GFS) [19] | Samples are gradually taken and concatenated from the geodesic path between source and target domains after PCA | Both small-dimension (PCA, Laplacian Eigenmaps) and high-dimension RKHS were experimented with. | Latent Discriminant Analysis (LDA), Kernel LDA (KLDA), Partial Least Squares (PLS) |
| | | CV | Geodesic Flow Kernel (GFK) [20] | Kernel-based method. Source and target datasets are embedded into the Grassmann manifold. The authors further propose a geodesic flow kernel, equal to integrating over all the Geodesic Flow subspaces. | The subspace dimension is small (PCA based). However, infinite subspaces are used. | Rank of Domain (ROD) [20], based on KL divergence |
| 3 | **Feature space alignment**: Learns a *linear* mapping function to align source subspace to target subspace. | CV | Subspace Alignment [21] | Aligning source and target subspaces using a mapping function (covariance matrix between source and target eigenvectors) and then calculating a similarity function to compare data. | The subspace dimension is small (PCA based). | Bregman matrix divergence |
| 4 | **Feature Transformation:** Typically involves *non-linear* transformation to either source or target or both; examples include reconstruction and transformation under constraints. | CV | Transfer Sparse Coding (TSC) [22] | Learning sparse representation (non-linear transformation) and then bringing the transformation closer by using MMD. | The subspace dimension is sparse (greater than source/target feature dimensionality) | Maximum Mean Discrepancy (MMD) [14] |

**TABLE 6.** Instance re-weighting based Shallow DA approaches.

| Sr. No. | Approach | Data Domain | Example | Key Idea | The criterion (/criteria) for distribution difference / Discriminative Methods |
|---|---|---|---|---|---|
| 1 | **Soft Selection**: The aim is distribution difference reduction by reweighting the source data and then doing a task (typically classification) on the reweighted source data. | NLP | Transfer Adaptive Boosting (TrAdaBoost) [25] | Iteratively re-weighting is done by increasing the weights of misclassified target domain samples and decreasing the weights of misclassified source domain samples | SVM / derivative of SVM – TSVM |
| 2 | **Hard Selection**: It is typically a two-step exercise. Firstly, an instance re-weighting exercise is done, and then based on a threshold (which reflects how close are two domains are), instances below the threshold are rejected, and only the ones above are included. | NLP | Knowledge Adaptation [26] | Maximum Cluster Difference (MCD) is defined as the threshold metric between the centroid cluster of instances from the source and target domain. | Maximum Cluster Difference (MCD) [26] |
| 3 | **Density ratio:** Weights an instance based on the likelihood of it coming source or target domain. These are direct techniques | Others – Structured Data | Independent Estimation [27] | Estimating domain likelihood based on classifier and selecting samples thereof. | Naïve Bayes, Logistic Regression, Decision Tree and SVM |
| | | CV, Time Series | Kullback-Leibler Importance Estimation Procedure (KLIEP) [28] | Estimating domain likelihood based on minimization of KL Divergence, selecting those samples which have less KL divergence | KL Divergence |

of each view simultaneously, assuming the features are not (much) common (i.e., $\chi^{viewi} \neq \chi^{viewj}$), very similar to heterogenous DA where $\chi^{source} \neq \chi^{target}$. Also, similar is Domain Separation Networks (DSN) [31], where private and shared feature spaces are orthogonal, as far as possible, the endeavor is to have (in different co-training strategies) features split into two mutually exclusive views. Blum & Mitchell, in their co-training strategy [32], solved the NLP text classification problem, using as one of the views the anchor texts of hyperlinks of pages pointing to the page and another view as the text of the page. The features are taken to be dissimilar. Due to reduction of the bias of predictions on unlabeled data, Ruder [33] mentions that Tri-training is one of the best multi-view training methods.

## IV. DOMAIN ADAPTATION IN DEEP LEARNING

Since deep neural networks are associated with high accuracy (or any required metrics) and can provide state-of-the-art (SOTA) results, there has been increased usage of deep neural networks in many AI and ML applications and tasks. However, these networks also face domain shift problems and are not able to adapt to different (from source domain) data distributions and provide the same SOTA results. Further, given that deep neural networks require a large amount of labeled data to train and the availability of labeled data is a concern (it is costly, arduous, or at times infeasible), it is much required that DA is supported for deep neural networks. Unlike DA in shallow learning, the focus of DA in deep learning is to include DA in the deep learning process and pipeline such that transferable representations are learned. In this direction, the earliest work, Glorot et al. [34], included Stacked Denoising Autoencoders (SDA) on amazon.com product reviews to do sentiment analysis for different products. After that, substantial work has been done in the CV area, with NLP picking up (again) fast in the recent past – primarily due to the availability of transfer learning in NLP using transformers and attention architectures. DA research has now gathered pace to solve real-world problems (like multi-modal data support, data restrictions, and scarcity).

Table 7 lists the Deep DA methods and approaches and further extends on the deep DA categorization mentioned by Wang and Deng [35]. However, [35] only focused on Deep DA techniques for the visual domain. In contrast, we aim to include more deep DA approaches, which are data domain-specific and review progress on other existing approaches. Also, our emphasis is to learn more about DA in unsupervised settings; supervised settings, semi-supervised and pseudo-semi-supervised are included for completeness or novelty.

### A. DISCREPANCY-BASED METHODS

These methods build on the shallow domain adaptation methods, map the features to a high dimensional RKHS space, and

**TABLE 7.** Deep DA methods and approaches classification.

| S. No. | Approaches / Methods | Key Idea | Sub Classification |
|---|---|---|---|
| 1 | Discrepancy-based methods | These methods look to reduce domain shift by fine-tuning the deep network<br>Initially an extension of shallow learning with some new techniques and discrepancy distance metrics. | • **Metrics-based:** Summary statistics are extracted from $P(X^s)$, $P(X^t)$ and a representation is built, such that some kind of statistics of $P(X^s)$, $P(X^t)$ is matched/aligned.<br>• **Architecture-based:** Deep architectures are built such that the transfer is easy. Methods also include modifying the architecture so that it is able to accommodate the domain shift better. |
| 2 | Adversarial methods | In this, two networks compete or are in an adversarial relationship to understand domain invariant features or similar aspects. Typically, a domain discriminator and an adversarial objective are used towards domain confusion. | • **Adversarial discriminative models or methods**<br>• **Adversarial generative models or methods**<br>• **Adversarial reconstruction-based models or methods:** These are also referred to as domain mapping (e.g., Image to Image translation); these aim to reconstruct one domain content in another domain style. |
| 3 | Multi-Domain Approaches | Deep DA in multi-source or multi-target domain settings differs from the above-mentioned techniques. The aim is to gather if the multiplicity of domains in multi-source helps in a more robust model, and multi-target helps the model to generalize better. | • **Multi-Source DA methods**<br>• **Multi-Target DA methods** |
| 4 | Hybrid Approaches | An amalgamation of two or more strategies mentioned before. | |
| 5 | Specific Data domain methods | Approaches that are relevant for a specific domain and not typically are not data domain agnostic. These also include DA methods for real-world challenges, practical settings, emerging methods, and techniques | • **Multi-modal Data DA methods**<br>  ○ Homogeneous Multi-modal DA<br>  ○ Heterogeneous Multi-modal DA<br>• **Data Domain Specific Methods**<br>  ○ Natural Language Processing (NLP) specific methods<br>  ○ Computer Vision (CV) specific methods<br>  ○ Speech specific methods<br>  ○ Time-series specific methods |
| 6 | Emerging DA Methods | Emerging DA methods for practical settings and real-world challenges | |

understand the discrepancy using metrics like MMD or similar. The difference being distribution difference is understood and aligned using deep features against the hand-crafted features of shallow DA methods.

Figure 4 shows the typical structure/architecture of networks implementing discrepancy-based methods – discrepancy metrics or representation of the network (along with discrepancy metric) or loss is used to regularize the network. Domain adaptation can happen at single or multiple layers (called adaptation layers in Figure 4)

### 1) DISCREPANCY METHODS: METRICS-BASED

Deep Domain Confusion (DDC) [45] was the first key idea that jointly optimized task (classification) and domain confusion. Similar to Figure 4, DDC used 2 parallel networks with one network as supervised (classification loss was included) and the other network as not supervised. Domain (confusion) loss is used to adapt two fully connected layers with the idea: features that the network learns should be agnostic, i.e., they should lie in a feature space where domain information is lost while the class information is intact. MMD [14] is used as a discrepancy metric for domain loss. Extending (2), DDC mentioned the joint loss for domain adaptation as

$$\mathcal{L}_{Domain\_Adaptation} = \mathcal{L}_{Classification} + \lambda MMD^2(P(X^s), P(X^t)) \tag{4}$$

Equation (4) also helps us understand that the discrepancy metric (MMD or similar) acts like a regularizer for the overall network. Later on, many works have built up on the DDC's key idea by using different/similar discrepancy metrics. Discrepancy metrics often used in deep DA are mentioned in Table 8.

The work of Kashyap et al. [46] further segregates the divergences into 3 classes – Geometric (distance between vectors), Information-theoretic (distance between probability distribution), and higher-order measures (amongst higher

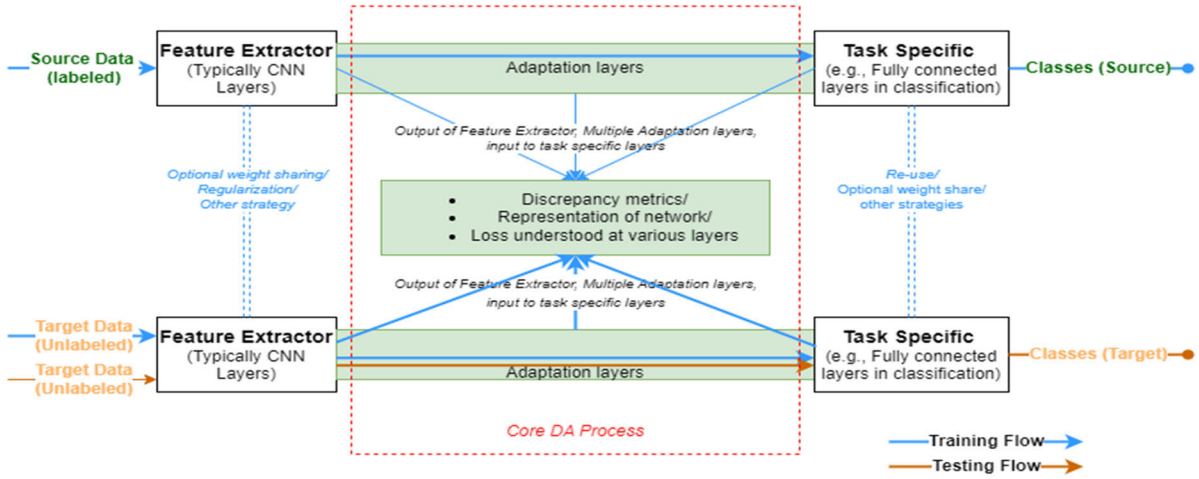


**FIGURE 4.** Typical network structure/architecture of deep neural networks implementing discrepancy-based methods for domain adaptation. The feature extraction layers are shared or regularized. The layers after feature called "Adaptation layers" include the discrepancy metrics or a representation of network along with the discrepancy metric. Task specific layers do not take much part in adapting the features of source and target. Best viewed in color.

**TABLE 8.** Discrepancy metrics and usage.

| Sr. No. | Metric | Data Domain | Usage/remarks |
|---|---|---|---|
| 1 | Maximum mean discrepancy (MMD) [14] | NLP, CV, Time Series, Multimodal | Distance between mean embeddings of two features (one from source and one from target), or simply the distance between means of two distributions. This metric is widely used in DA. |
| 2 | DeepCORAL [36] | Primarily CV | Extends the use of CORrelation Alignment (CORAL) [37], which uses covariances (second-order statistics) of source and target features in Deep Neural Networks. Other variants (LogCORAL, LogD-CORAL) of these metrics are also used in unsupervised DA. |
| 3 | Contrastive domain discrepancy (CCD) [38] | CV | Use of conditional distributions to incorporate label information in Contrastive Adaptation Networks (CANs) [38] |
| 4 | MK-MMD (multiple kernel – MMD) [39] | CV | Multiple Kernel Learning variant based on MMD, first used in Deep Adaptation Network (DAN) [39] Instead of matching using the first moment of data, it uses the second moment, i.e., the Multi Kernel variant of MMD (MK-MMD). We do not have one RKHS here but multiple RKHS, and we have multiple kernels of multiple RKHS. In a neural network, this is done by doing multiple kernels on multiple feature spaces (output of fully connected layers) |
| 5 | Joint maximum mean discrepancy (JMMD) [40] | CV | Instead of marginal distribution – joint distribution is used. It takes MK-MMD to the next level, where it is jointly minimized, i.e., MMD between representations of multiple layers / multiple kernels together in a joint way. So we don't minimize MMD independently but jointly over representations of multiple layers |
| 6 | DeepJDOT [41] | CV | Adapted in deep neural networks from Joint distribution optimal transport (JDOT) [42], which estimates a non-linear transformation between joint features/labels of two domains. The discrepancy metric in JDOT is the 1-Wasserstein metric. Also, one can show that kernelized MMD is equal to Wasserstein distance. |
| 7 | Kullback-Leibler (KL) divergence [43] | NLP, CV, Time Series, Multimodal | [43] used KL divergence to understand the distribution distance in the embedding layer between the source and target domains. |
| 8 | Sliced Wasserstein discrepancy [44] | CV | [44] used optimal transport theory (i.e., a variation of Wasserstein distance) to align feature distribution between domains and task-specific boundary |

moment distance between distributions or distance between projections or distance between representations).

### 2) DISCREPANCY METHODS: ARCHITECTURE-BASED

In these methods, the focus is more on learning more transferable features and architecture than the metric. The underlying principle with architecture-based discrepancy methods is that information about the domain change (source to target) is only an affine transformation away, i.e., there exists a small transformation on weights that can help the transformation from source to target features. This small transformation can be the affine transformation or Multi-Layer Perceptron (MLP) / Deep Networks themselves.

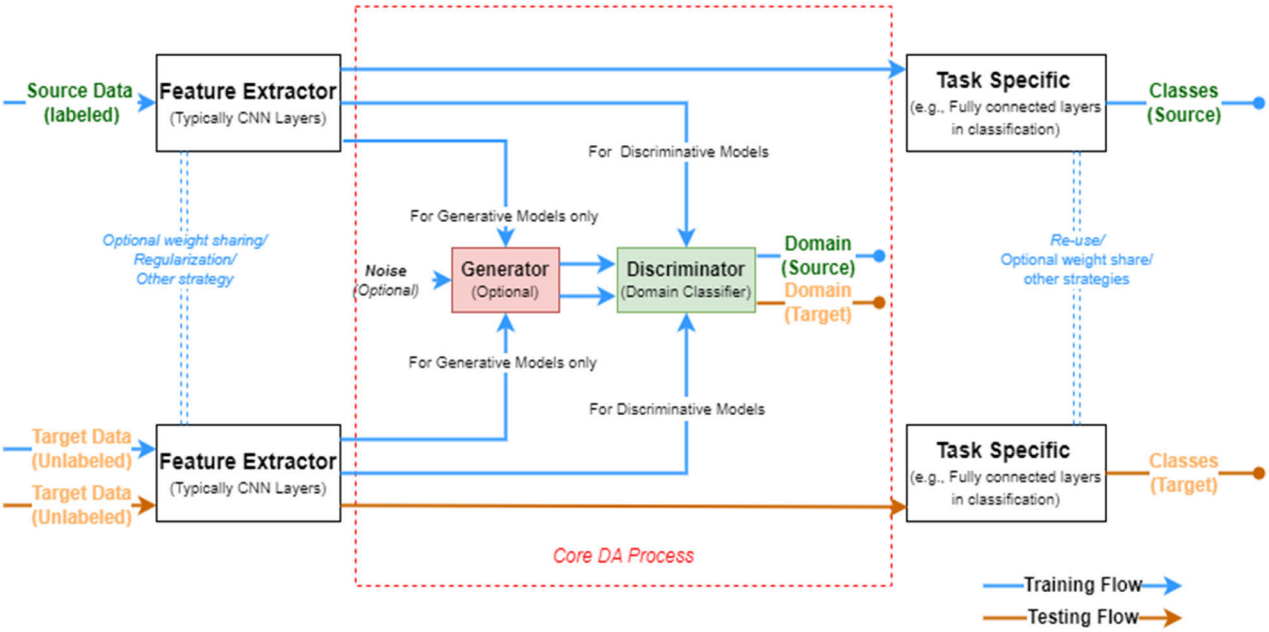


**FIGURE 5.** Typical network structure/architecture of deep neural networks implementing adversarial-based methods for domain adaptation. The generator(s) are optional – used only to create synthetic data if required. Domain Discriminator looks to discriminate (classify) source and target domains, while task-specific (say classification) is done by task-specific network. Best viewed in color.

Deep Adaptation Network (DAN) [40] uses the concept that in convolutional deep networks transition, earlier layers understand generic features while the later layer understands task-specific features. They froze the initial layers, fine-tuned the middle layers, and looked at discrepancy-based methods like MK-MMD (multiple kernel – MMD), a variant of MMD [13], to adapt later layers. Typically, the discrepancy-based methods look to align the marginal distribution of source and target data, but there are different approaches, too, like Joint Adaptation Network (JAN) [41]. JAN further improved on DAN architecture by learning joint distributions of multiple domain-specific layers across domains and using the joint maximum mean discrepancy (JMMD) criterion; they used a representation ($\varphi$) of the network itself.

Similarly, JAN-A [41] builds further on JAN architecture that there is now another network ($\theta$) that computes representation on top of network representations ($\varphi$). This not only minimizes the JMMD but also learns the network ($\theta$) – maximum is over the $\theta$ network, and the minimum is over the JMMD – an adversarial objective (min-max). Computer Vision (CV) also uses normalization layers as a key architectural concept for DA. Given that this is specific to CV, it is detailed in the normalization layers (refer to section Normalization layers). The hypothesis behind this is batch normalization (BN) layer represents domain-related knowledge. Transferrable Prototypical Networks (TPN) [47] focus on discrepancy (distances) for each class in an embedding space of 3 datasets - source only, target only and a mix of source and target. It also assigns ''pseudo-labels'' to unlabeled target samples. Adaptation is done so that the prototype of each class is close in the embedding space.

### B. ADVERSARIAL METHODS

The idea behind adversarial set of methods is to enhance domain confusion while still being robustly trained to understand domain segregation (adversarial objective). This is closely related to Generative Adversarial Network or GANs [4], which includes two networks – Generator and Discriminator – in an adversarial setting. The generator aims to produce output (typically images) to fool or confuse the discriminator, while the discriminator, on the other hand, tries to segregate it into real and fake. In DA, the idea borrowed is that discriminator should be able to segregate the domain distribution of source and target domains (say by using domain invariant features). Adversarial Discriminative Domain Adaptation, or ADDA [48] introduced a generic framework (similar to Figure 5) for DA using adversarial models. Typical adversarial discriminative architectures follow a Siamese architecture with source and target stream and are trained on task loss (typically classification) and either an adversarial loss or a discrepancy loss. In contrast, adversarial generative architecture (in its simplest form) includes a generator that generates other domain (typically target) mapping from the first domain (typically source); after that, the generated mapping and other mappings then follow adversarial discriminative architecture.

#### 1) ADVERSARIAL DISCRIMINATIVE MODELS

One of the seminal works in deep DA is Domain-Adversarial Neural Network (DANN) [49] (refer to Figure 6), which supports the idea of adversarial domain adaptation, i.e., learning task should be discriminative yet, it should encourage domain confusion. They showed that any feed-forward model

could support adaptation if augmented with a novel gradient reversal layer.

DANN is the most widely used DA approach across all data domains. In CV, DANN was initially used for digit recognition and image classification. Later on, DANN or its derivatives are also used for more complex tasks like semantic segmentation and object detection. In the case of semantic segmentation, a Siamese network (consisting of two parallel tracks) approach is taken, where one track processes source samples and the other track processes target samples. Due to the inherent complexity of tasks – Domain alignment (Domain Classifier – pink network in Figure 6) is present at various layers/stages, and the convolution layers (input to feature extractors) and deconvolution layers (feature extractors to semantic map) are aligned (shared, mapped or statistical metric is used). Hoffman et al. [50] used 2 more losses other than the regular semantic loss – one loss to adapt to category-specific parameters, i.e., category-specific adaptation and the other loss to reduce ''global distribution distance,'' i.e., global domain alignment. Huang et al. [51] looked at aligning features at each layer of the network.

In NLP, DANN has been widely used for classification tasks – Text Classification ([52], [53]) and Sentiment Analysis ([49], [54], [55]). DANN or its variants are also used for Named Entity Recognition (NER) ([52], [56]) and Parts of Speech (PoS) Tagging [57]– structural prediction tasks.

Adversarial Discriminative Domain Adaptation or ADDA [48] model also uses similar philosophy as DANN but differs in that feature extractors are not shared between source and target, and the loss function that is used in ADDA is GAN loss while DANN uses min-max loss and the training is multistep. Conditional Domain Adversarial Networks (CDAN) [58] use a conditional discriminator, taking input from both feature extractor and classifier. Work of Shen et al. [54], instead of using a pure classifier in the discriminator, used the loss as Wasserstein distance (similar to Wasserstein GAN by Arjovsky et al. [59]) during training between source and target samples. Inspired by the multi-view strategy, Du et al. [60] proposed Dual Adversarial Domain Adaptation (DADA), having two ''joint'' discriminators, supporting all the classes of source and the target domain ($2K$-dimension), pitted against each other and back-propagating into feature extractor. They also used a source class predictor to classify source labels and provide pseudo labels. The latest attempt to improve adversarial discriminative models is Smooth Domain Adversarial Training (SDAT) [61], which mentions that reaching smooth minima only for the task-specific loss (and not the domain discriminator loss) helps better adapting to the target domain.

#### 2) ADVERSARIAL GENERATIVE MODELS

Adversarial generative models are different from Adversarial discriminative models in that they have a generative component (typically, a generator of GAN) along with the discriminative component of discriminative models. This generative component typically creates synthetic target data from labeled source data. This synthetic labeled target data alleviates the need for labeled examples in target domains. Then the network is trained to assume there is no or little domain shift present in the synthetic data. The source mapping component is the generator that maps the source domain into the target domain. Therefore, colloquially these generators are also known as domain mappers. One of the earliest works in adversarial generative models is Coupled Generative Adversarial Network – CoGAN [60]. As the name suggests, two GANs run parallel, and weight sharing happens in the initial layers for generators and the final layers of discriminators. These layers capture high-level features in discriminators and high-level semantics in generators. This helps the GAN to understand the joint distribution of domains. In CoGAN, the target domain is transformed into the source domain, and then the classification happens.

Typically, the DA is specific to a task (shared across two domains); however, PixelDA [61] used an adversarial generative DA setup to provide a framework that is decoupled from task-related aspects. Typically, source images are transformed into target-like images; however, Generate-to-adapt [62] uses GANs for domain adaption with Generator creating source-like images for target domain cases. It uses the embeddings (learned during training) of images as the latent space as an auxiliary input to the GAN to create source-like images from the generator and discriminator, discriminating the domain (real/fake) and providing class labels. Other examples of adversarial generative models in the speech domain are Park et al. [63] and Augmented Cyclic Adversarial Learning (ACAL) [64].

#### 3) ADVERSARIAL RECONSTRUCTION-BASED METHODS

Another variation of adversarial generative methods is reconstruction-based methods (on the same lines as shallow feature matching DA strategy): Reconstruction methods typically use Adversarial GAN-based networks or Autoencoder (AE) based networks to reconstruct one domain content in another domain style. Table 9 provides key ideas behind some Adversarial reconstruction-based methods. There are other methods in the literature that do not fully comply with the adversarial reconstruction definition but still are very close to its working.

- An example in NLP is AE-SCL: Ziser and Reichart [66] brought SCL [16] into the neural networks using Autoencoders; their network is called Autoencoder-SCL or AE-SCL. AE-SCL does not reconstruct the input but predicts if the pivot features will be present in the input or not. They used this for cross-domain sentiment analysis. They further improved AE-SCL using Pivot-Based Language Modeling (PBLM) [67] and Task Refinement Learning using PBLM (TRL-PBLM) [68].
- An example in CV is DiscoGAN: DiscoGAN [69] is also very similar to CycleGAN, the difference being that it does not have cyclic reconstruction loss.

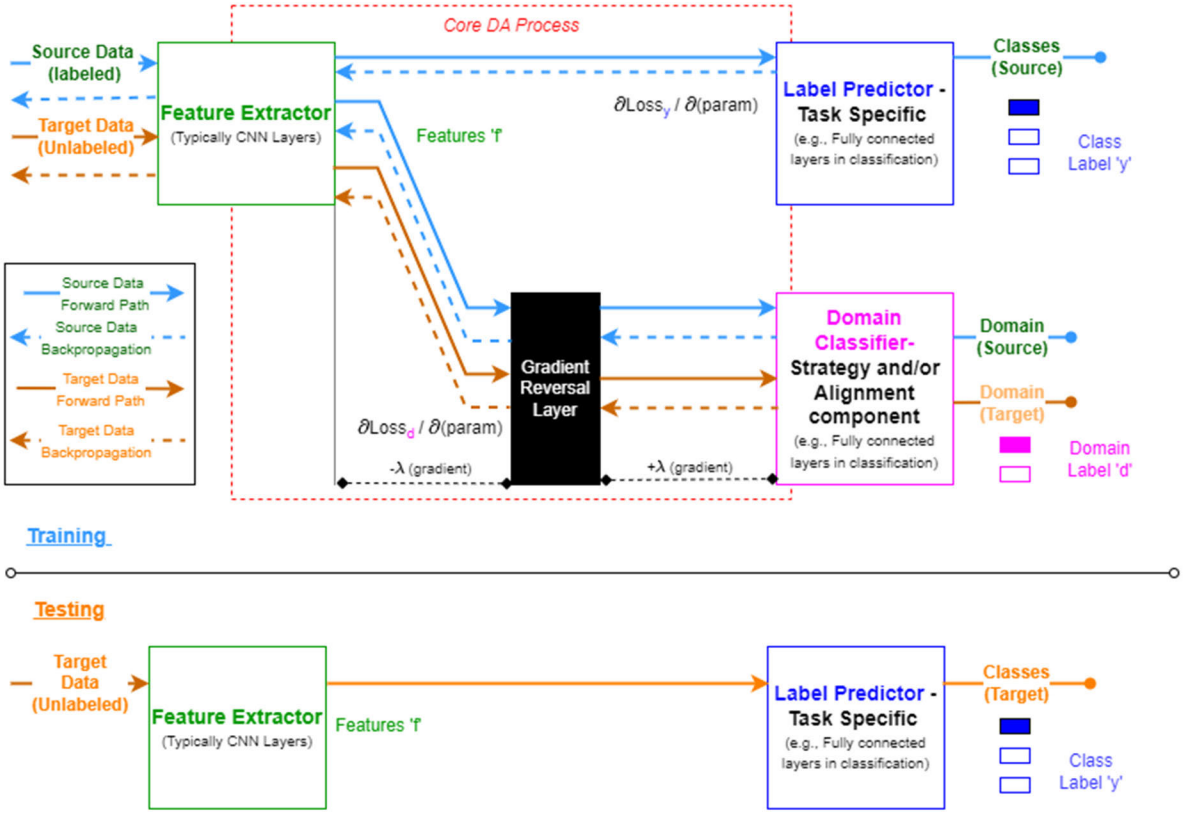


**FIGURE 6.** Domain-Adversarial Neural Network (DANN) trains two network together. DANN trains feature extractor (green network) and class/label predictor (blue network) on source data. DANN also trains feature extractor (green network) and domain classifier (pink network) on the source and target data. The gradient reversal layer (GRL) allows the feed-forward network to progress as it is; however, during the backpropagation, it changes (reverses but multiplies by a negative quantity) the gradient from domain discrimination, which leads to the feature extractor (green network) understands domain invariant features (domain confusion). λ helps to learn the classification features and then slowly learn domain features. Best viewed in color.

### C. MULTI-DOMAIN ADAPTATION

Multi-Domain DA setting differs from a typical DA setting in that either number of source domains would be multiple (called multi-source adaptation), or the number of target domains would be multiple (called multi-domain adaptation).

#### 1) MULTI-SOURCE ADAPTATION

To create more robust domain-adapted models, it makes sense to train the models on multiple sources. In earlier surveys (pre-deep learning), Sun et al. [70] mention training an individual classifier on individual source domains and a target domain and then merging the base classifiers or merging all the sources as one source and then training. For deep learning, Zhao et al. [71] again mention the use of discrepancy, adversarial, and feature alignment-based strategies. Another strategy or area explored is intermediate domain generation for adaptation; in this case, a domain is generated using domain generators (typically GAN-based).

Moment Matching for Multi-Source Domain Adaptation(M3DA) [72] created a multi-domain dataset called DomainNet with 6 domains. Further, they dynamically aligned moments of feature distributions of the multiple labeled source domain and the target domain. Zhao et al. [73] introduced Multi-source Adversarial Domain Aggregation Network or MADAN, which essentially uses CycleGAN (sub-domain aggregator discriminator for source domains and cross-domain cycle discriminator for source-target domains) coand creates a latent adapted domain for all source data and target data. Similarly, Russo, Tommasi, and Caputo [74] used CoGAN to adapt each source and target domain. Rebuffi et al. [75] used one residual adapters (which sit on the residual branch) for each domain. Yang and Hospedales [76] provided both multi-task and multi-domain perspectives using low-rank tensor methods; this work also provides an alternative to zero-shot learning.

In NLP, Guo et al. [77] introduce DistanceNet-Bandit, with distance metrics (DistanceNet) providing loss functions in addition to task loss along with using multi-armed bandit to control switching between multiple domains dynamically. Guo et al. [78] used meta-learning to combine predictors from each source-target domain.

In time-series, Zhu et al. [79] used a multi-adversarial strategy where multiple source domains (sample of roller bearings) were projected into a shared subspace, and domain

**TABLE 9.** Adversarial reconstruction-based methods.

| Sr. No. | Example | Data Domain | Key Idea |
|---|---|---|---|
| 1 | Domain Separation Networks (DSNs) [31] | CV | DSN is an Auto-Encoder (AE) framework that looks to understand features/representations private to the specific domain alongside shared features/representations. The shared and private spaces are orthogonal to each other; that way, DSN can separate domain-specific information from information shared across domains. Further, reconstruction loss (one amongst other losses) forces the network (decoder) to reconstruct the original input example using features present in that domain's private and shared domain features/representations. |
| 2 | Adversarial Adaptation of Synthetic or Stale Data [52] | NLP | The authors used the DSN concept from CV and applied it to NLP for text classification and named entity recognition (NER) tasks using single Bidirectional long short-term memory (BiLSTM). |
| 3 | Genre Separation Networks (GSNs) [62] | NLP | GSN used a variant of DSN called Genre Separation Networks (GSNs) in NLP, using deconvolution for reconstruction from word embeddings. |
| 4 | CycleGAN [63] | CV | CycleGAN is a GAN-based reconstruction-based method - a type of domain mapping (image-to-image translation) network which aims to reconstruct (based on cycle consistency loss) the source image back after translating it from source to target domain, alleviating the need for paired images. |
| 5 | Cycle-consistent adversarial domain adaptation (CyCADA) [64] | CV | CyCADA takes inspiration from CycleGAN to include cycle-consistent loss alongside other losses (image-level loss, feature-level loss, and task loss). All the losses together, when minimum, support DA. |
| 6 | Deep Reconstruction Classification Networks (DRCN) [65] | CV | DRCN used pair-wise reconstruction loss and focused on unsupervised reconstructing target domain data. DRCN can be visualized as a Convolutional Neural Network (CNN) having two pipelines – one for supervised classification of source domain labels, the other for reconstruction, sharing the same decoder (convolution and fully connected layer). |

invariant features were obtained. Xia et al. [80] introduced a moment matching-based intraclass multisource domain adaptation network, which measures the discrepancy (MMD) between each source domain and target domain samples.

#### 2) MULTI-TARGET ADAPTATION

Typically, DA follows the pairwise approach, with the source domain linked to the target domain. Inspired by [73], Gholami et al. [81] also look for shared information across domains. They propose Multi-Target DA-Information-Theoretic-Approach or MTDA-ITA, which uses private and shared spaces between source and target combination, much like Domain Separation Networks (DSN) [31]. Isobe et al. [82] used multi-target DA for semantic segmentation tasks using the individual source-target and individual bridges created amongst the pairs for collaboration. A student model is learned based on all the individual source-target model pairs using regularization on each individual source-target model pair. Similar knowledge distillation is understood in Multi-Teacher Multi-Target DA (MT-MTDA) [83]

### D. HYBRID METHODS

Hybrid methods indicate the amalgamation of multiple techniques discussed before for executing DA.

#### 1) ENSEMBLE-BASED METHODS

Ensemble methods contain multiple models, where the output of multiple models is combined, typically averaging in regression and voting in case of classification tasks. The diversity of the models makes sure that the deviation from correctness is not much. One of the most significant drawbacks of these models is that they are computationally expensive. Ensemble methods for DA can be segregated into two sub-techniques – pseudo labeling ensembling and self-ensembling.

In the case of the self-ensembling method, the combining of output is done on multiple outputs of a single model over time. Combining outputs over time is also known as temporal ensembling. French et al. [84] used Teacher Student (mean teacher variant) architecture proposed by Tarvainen and Valpola [85], as a self-ensemble technique for visual DA. The teacher network is first trained on the task and outputs floats (probabilities) instead of Boolean (0–1 integer) labels. The student then learns from the teacher, and the student can learn things better because the teacher informs the student of the nuances. Gradient descent is used to train the student network, while the exponential moving average of the student network is the weight of the teacher network. The training loss is a combination of a supervised and an unsupervised component. This architecture dramatically reduces the model parameters without compromising on accuracy metrics. In NLP, [86] also used adaptive ensembling, an extension to temporal ensembling, and classified political data while studying temporal and topic drift. They used a temporal curriculum and a student-teacher network.

Another data-centric variant of ensembling is pseudo-labeling ensembling, wherein the target domain labels are provided based on the combined perspective of comprising models. If most models converge, i.e., there is high

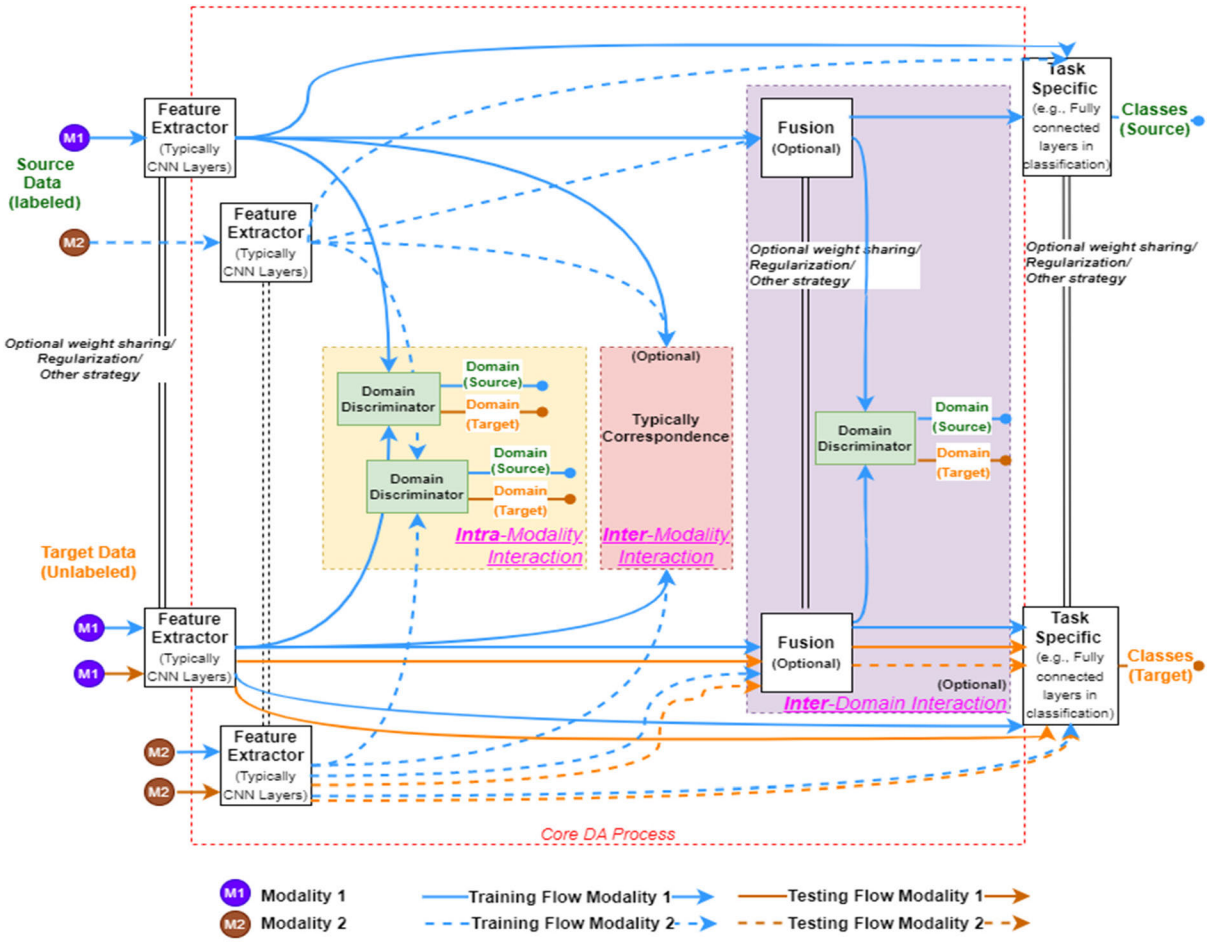


**FIGURE 7.** Typical network structure/architecture of deep neural networks implementing homogeneous multi-modal DA. Discriminators in Intra-Modality block force the feature extractors to understand domain-invariant features. Inter-Modality and Inter-Domain are optional and not seen in every multi-modal DA setting. Best viewed in color.

confidence in label class for a particular instance of the target domain. An instance of the target domain (not the source domain) is used for training the target classifier, hence the name of the technique pseudo-labeling. In computer vision, Saito et al. [87] proposed Asymmetric Tri-Training (ATT), which had two networks providing the labels for target domain instances – first trained on the source domain if the two networks converge, then the pseudo label is assigned to the target instance, and that data is used for training the third network. Final labels don't have to be provided at all times, and the probability score can also be used instead (examples: Zou et al. [88] and, to some extent, French et al. [84])

### E. MULTI-MODAL DOMAIN ADAPTATION

Multi-modal is a complex data domain with respect to DA, as the DA process has to take into account the different modality structures and different domain shifts (for each modality). In the case of heterogeneous multi-modal DA, the DA process must also take care of different feature spaces/ feature representations/ dimensions of feature spaces.

#### 1) HOMOGENEOUS MULTI-MODAL DOMAIN ADAPTATION

Most of the work in DA supports homogeneous data, i.e., feature space remains the same ($\chi^s = \chi^t$), but the shift is because of different data distributions, i.e., $P(X^s) \neq P(X^t)$. When both source and target domain would have at least two modalities, i.e., multi-modal, but still, the feature space (features fed for the task perspective) is the same, it is called Homogeneous multimodal DA. Typical homogeneous multi-modal architecture (refer to Figure 7) does implement intra-modality interaction compulsorily; however, it is seen that implementation of inter-modality and inter-domain aspects is optional.

Qi et al. [89] created a multi-modal DA network with attention and fusion modules along with hybrid domain constraints to learn domain invariant features. The *intra* and *inter* units in the attention module help to understand the relationship among modalities. The bilinear model approach ([90], [91]) was used for fusion, and then tucker decomposition was used to support computational (GPU) [92] restriction.

For social media event rumor detection, Zhang et al. [93] proposed Multi-modal Disentangled Domain Adaption

(MDDA), which looks to resolve two challenges – entanglement and domain. Disentanglement of event content with rumor style was done as part of the first challenge, and domain shift was tackled in the latter challenge (with only rumor style taken after the first challenge). The network learned only a transferrable rumor style with the alignment of feature distributions over different events.

Multi-Modal Self Supervised Adversarial Domain Adaptation or MM-SADA [94] uses two modalities – optical flow and RGB of EPIC-Kitchens video dataset, and understand if the fine-grained action recognition (depends highly on the environment) can be improved across dataset domains. They used self-supervision across two domains with both modalities and adversarial adaptation between each modality of source and target data (i.e., one discriminator for RGB and one for optical flow).

Li et al. [95] look at DA amongst multiple modalities from domains (scripted source, improvised source). They use an emotion recognition model based on adversarial training (which helps to remove domain difference between emotion elicitation approaches) and a soft label loss approach (which helps to understand non-rigid emotions and to consider emotion and domain categories simultaneously).

#### 2) HETEROGENEOUS MULTI-MODAL DOMAIN ADAPTATION

One of the most prevalent real-world data is the heterogeneous multi-modal domain; as deep networks look to use more heterogeneous multi-modal data, it is imperative to learn DA in heterogeneous multi-modal settings. The DA, in the case of heterogeneous data, is carried out by extracting features of two domains using separate network and the task level aspects either by sharing weights (strong parameter sharing) or weakly parameter-shared weights as in the work of Shu et al. [96].

The importance of heterogeneous multi-modal DA lies when one of the modalities is missing in the target domain: consider the source domain having modalities *m1* and *m2*, while the target domain may just contain *m3* with missing *m4*. Ding et al. [97] look at solving a real-world 'Missing Modality Problem' by introducing Missing Modality Transfer Learning via latent low-rank constraint (M2TL). The transfer of learning is twofold – one, from one database to another (cross-database transfer), and two, from source modality to target modality (cross-modality transfer). They use low-rank matrix constraint to learn subspace within a database across modalities and MMD to couple databases in the source domain (known modalities).

Conditional adversarial domain adaptation [58] uses conditional domain adversarial networks (CDAN), a variant of the adversarial discriminative model, which assists adversarial adaptation by employing discriminative information understood in the classifier predictions. The discriminator is conditioned on the cross-covariance of domain-specific feature representations and classifier predictions. CDAN can adapt to multi-modal data distributions and can support scenarios involving higher-dimension also (supported by a variant called Randomized Multilinear (RM) conditioning).

Athanasiadis et al. [98] present Domain Adaptation Conditional Semi-Supervised Generative Adversarial Networks (dacssGAN) in the realm of emotion recognition, where domains (audio, video) are heterogeneous and multi-modal. The network uses GANs and conformal prediction techniques [99] to implement DA.

Seo et al. [100] aim to improve audio-visual sentiment analysis performance using text modality during the training phase by ''transferring knowledge'' of unimodal (text modality) to other modalities (audio and visual). The knowledge transfer employs the reduction of distribution differences of feature representation in data for each modality.

In NLP, Cross-lingual translation also falls under heterogeneous tasks as the words and the construct of the two languages are very different, leading to an assumption that input features don't match. i.e., $\chi^s \neq \chi^t$. Various attempts, including Conneau et al. [101], have been made to support cross-lingual DA as an unsupervised task; however, Søgaard et al. [102] showed that the underlying assumption that the words are isomorphic in a language is incorrect. They further suggested that a weakly supervised solution outperforms (the metric used was bilingual dictionary induction scores) unsupervised cross-lingual DA. Conneau et al. [103] mentioned that pre-trained models (discussed later in the section Pre-Trained Models) achieve better results in unsupervised cross-learning representation translation tasks. Generative adversarial text-to-image synthesis [104] provided a way to generate an image based on text, translating visual concepts of pixels from characters using a convolutional-recurrent neural network. Along similar lines, StackGAN [105] also created photo-realistic images in two stacked steps from the text.

### F. DOMAIN ADAPTATION IN COMPUTER VISION (CV)

This section focuses on DA strategies typically only seen in the computer vision data domain and not shared with other data domains.

#### 1) NORMALIZATION LAYERS

Normalization layers help maintain a stable training of neural networks and are used in nearly all neural networks. A few examples of normalization layers in regular neural networks are batch normalization or batchnorm [3], layer normalization or layernorm [106], instance normalization or instancenorm [2], and group normalization or groupnorm [107].

Chang [108] created a DA framework using a domain-specific batch normalization layer – other model parameters were shared between domains. Li et al. [109] proposed the Adaptive Batch Normalization (AdaBN) layer. The intuition behind the layers is that these layers learn domain knowledge in contrast to weights learning task knowledge and biases learning some sort of priors. Carlucci et al. [110] in AutoDIAL built further on [109] AdaBN layers and used DA

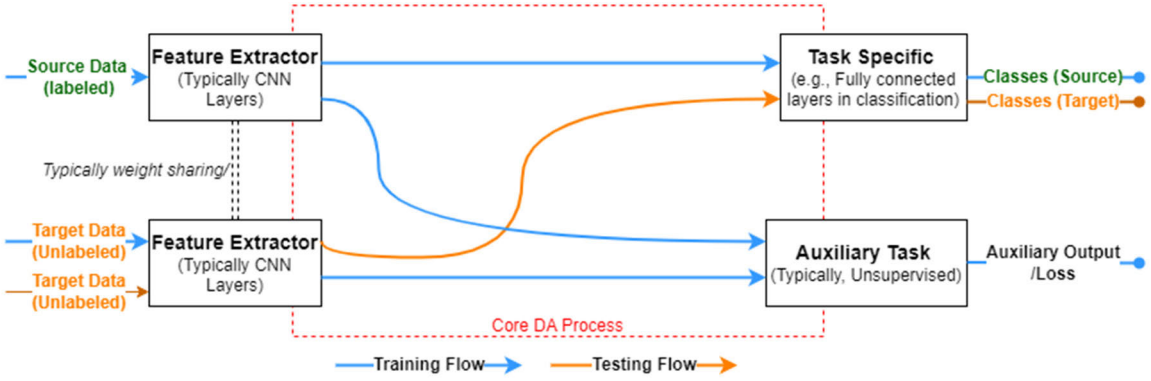


**FIGURE 8.** Typical self-supervision network structure. It is a multi-task network and includes an auxiliary task which aims at understand the feature distribution mode, but does not impact the core DA task but "provides" knowledge of sorts to core DA task. Best viewed in color.

layers amongst the standard CNN layers. The purpose of these layers was to normalize the target and source mini-batches (separate for two domains) but influenced by each other based on a parameter learned as part of the training process.

Roy et al. [111] proposed Domain-specific Whitening Transform (DWT) – domain alignment layers to compute intermediate feature covariance matrices, along with Min-Entropy Consensus (MEC) loss (a merger of entropy and consistency loss) for coherent predictions for sample.

#### 2) SELF -SUPERVISION METHODS

Self-supervision DA methods look at joint training of an auxiliary self-supervision task alongside the main task and therefore are also aligned to multi-task. In the Deep Reconstruction Classification Network (DRCN), [65] had a deconvolution network to reconstruct the image (an auxiliary self-supervised task) while the convolution network performed the label prediction (main task). The feature mapping parameters were shared in DRCN, very much similar to Figure 8. The intuition is that the main task receives knowledge transfer from the auxiliary task.

Carlucci et al. [112] used the auxiliary task of jigsaw puzzle solving (permutation index) while solving the main task as a DA/DG strategy. It is noted that typically the auxiliary task is an unsupervised task; however, the main task is a supervised task. Xu et al. [113] further increased the number of auxiliary tasks (image rotation prediction, flip prediction, and patch location prediction), further underlying those low-level differences (like pixel-level reconstruction/prediction) are not much useful in DA. In contrast, high-level structural task (like part of image rotation) is very useful. Kim et al. [114] showed that the self-supervision technique is useful even with few labeled instances in the source domain. They used within-domain instance discrimination (in-domain self-supervision) and cross-domain matching (across-domain self-supervision) to learn features that are domain-invariant as well as discriminative.

### G. DOMAIN ADAPTATION IN NATURAL LANGUAGE PROCESSING (NLP)

This section focuses on DA strategies typically only seen in the NLP data domain and not shared with other data domains. Most of the work in DA has been done in the CV area, though the origins of DA have been in NLP. For example, DANN [49] was initially applied to sentiment classification, but later it was used for computer vision classification tasks. Ramponi and Plank [115] categorize NLP domain adaptation models into Model-Centric, Data-Centric, and Hybrid. Model-Centric models (focus on augmenting the feature space, tinkering with loss functions, and changing the architecture of the model), discussed before, has been used in other applications and computer too. Pre-trained models are Data-Centric models and are discussed below, and hybrid models are discussed in the section Hybrid Methods.

#### 1) PRE-TRAINED MODELS

The Data-Centric models are not shared with computer vision tasks, perhaps because these models focus on data elements, different in computer vision and NLP, to support adaptation. These models are less prevalent but, of late, have picked up the interest of researchers. BERT Devlin et al. [116] was a model to revolutionize transfer learning– other methods include pseudo labeling, pre-training (zero-shot) (example: Multilingual BERT)/fine-turning (including multi-phase) (example: SciBERT [117] / BioBERT [118]).

Figure 9 provides a typical pre-trained training strategy, and Table 10 lists different pre-trained training data and strategies. Based on the DA definition, Pre-training and fine-tuning are not kinds of DA processes, but these transformer-based language models are task agnostic in the sense that they can be fine-tuned on specific tasks using a small dataset. It is included in this survey for completeness.

AdaptaBERT [119] used a two-step approach for domain-adaptive fine-tuning. In the first step, they performed domain tuning by taking contextualized word embeddings (unlabeled source and target domain data) and maximizing the

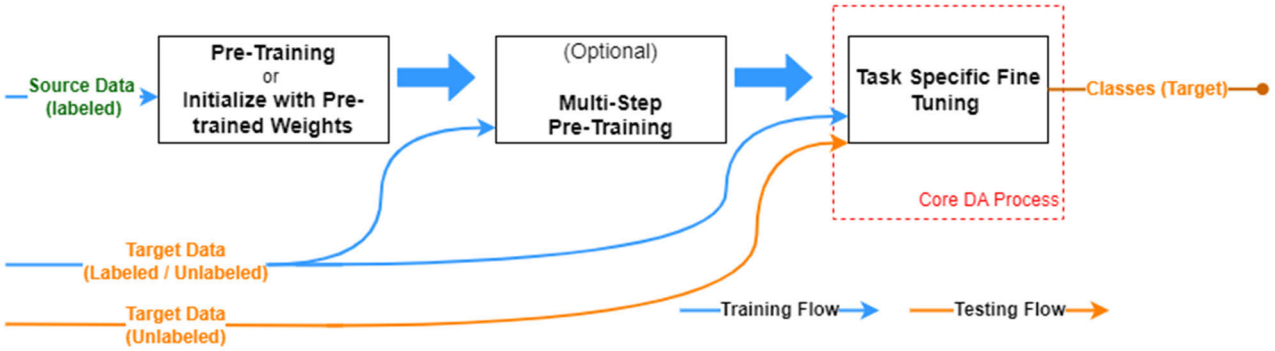


**FIGURE 9.** Typical Pre-trained Training strategy. Pre-training is typically task agnostic; future steps are required to adapt the model to the task in question. An optional multi-step pre-training is done to reduce the data distribution gap of source data and target data. Best viewed in color.

**TABLE 10.** Pre-trained training data and strategies.

| Sr. No. | Reference | Overall Strategy | Pre-training (Single or Multiple) | | Fine Tuning | |
|---|---|---|---|---|---|---|
| | | | Data | Strategy | Data | Strategy |
| 1 | BERT [116] | Pretraining only | General Domain (English Wikipedia and BookCorpus) | Masked LM | Domain-Specific | Task-Specific |
| 2 | Multilingual BERT (M-BERT) | Pretraining only | Text from 104 languages (Cross-lingual Natural Language Inference (XNLI [120]) dataset) | Masked LM | Domain-Specific | Task-Specific |
| 3 | ULMFIt [121] | Multi-phase Adaptive Pre-training | General Domain (Wikitext -103)<br>Domain-Specific | AWD-LSTM LM [122]<br>Discriminative fine-tuning [123] and slanted triangular learning rates | Domain-Specific | Task-Specific with Gradual Unfreezing (e.g., Back Propagation Through Time for Text Classification (BPT3C) |
| 4 | BioBERT [118] | Multi-phase Adaptive Pre-training | Specific Domain corpora (PMC full-text articles and PubMed abstracts)<br>Task-Specific Data sets | Initial weights of BERT<br>Named Entity Recognition, Relation Extraction, and Question Answering | Domain-Specific | Task-Specific |
| 5 | DAPT + TAPT [124] | Multi-phase Adaptive Pre-training | Broader Corpus from Domain (biomedical (BIOMED) papers, computer science (CS) papers, newstext from REALNEWS, and AMAZON reviews)<br>Unlabeled data closer and relevant to task distribution (total 8, 2 data sets for every four domains) | Domain Adaptive Pre-Training (DAPT) – Masked LM Loss on ROBERTa [125]<br>Task Adaptive Pre-Training (TAPT) | Domain-Specific | Task-Specific |
| 6 | STILTS [126] | Auxiliary Task Adaptive Pre-training | Unlabeled data (not very different from task data)<br>Labeled-data | LM on unlabeled data (BERT [116], ELMo [127], GPT [128])<br>Intermediate Task Training (target or auxiliary task) | Domain-Specific | Task-Specific |

probability of masked tokens. In the second step, they focused on task tuning by taking labeled source data and back-propagating for the desired task (PoS tags in this case).

### 2) MULTI-VIEW LEARNING

Another NLP-specific DA technique is multi-view training (also discussed briefly in heterogenous DA). Different views of data are used to train different models in multi-view training. The views differ from each other in the following dimensions (or a combination of dimensions):

1. Architecture of models
2. Features
3. Data used for training

The philosophy behind multi-view training is that the views complement each other, and the collaborated models improve each other's performance. Examples of multi-view training are Co-Training [31], Democratic Co-Training [129], and Tri-Training [130]

### H. DOMAIN ADAPTATION IN SPEECH

In speech domain adaptation tasks, the focus is to first identify which elements of the data are actually speech and not noise; for the elements identified as speech, then the focus is either recognition of speech called Automatic Speech Recognition (ASR) or adapting to a speaker. Text-to-speech (TTS) is a multi-modal variety where the output modality (space) is speech. The DA strategies that are typically employed are discrepancy based ([131])(refer to section Discrepancy-Based Methods), adversarial-based ([132], [133]) (refer to section Adversarial Methods), pseudo-semi-supervised training based ([131]) (refer to section Pseudo-Semi-Supervised Domain Adaptation) and knowledge distillation based ([134], [135]) (Ensemble-based methods or Teacher-student based, refer to section Ensemble-Based Methods).

One speech-specific strategy understood is the work by Zhang [136], where a pretraining process is undertaken on the DNN model using unlabeled target domain data first. Later, labeled source data is used to fine-tune the network. The intuition behind the pretraining process is to seek shared representation.

### I. DOMAIN ADAPTATION IN TIME-SERIES

Typically, the tasks that are prevalent in time-series DA are classification (generally 2 class classification) and forecasting (predicting based on past time-stamped information). Further, the problems solved are univariate and multivariate, i.e., involving multiple time-stamped variables used for prediction, e.g., pressure, temperature and flow rate predicting fault in a power station. Jin et al. mention [137] the complexity in time-series DA as two-fold:

1) **Varying*input and output space:*** The *output space* of the source domain time-series (say, the flow rate in the power station) may be different from the *output space* of the target domain time-series (say, a count of units in a warehouse). Hence, it is imperative that not only domain-invariant features are captured but also domain-specific features be captured as in Domain Adaptation Forecaster (DAF) [137]. Similarly, *input space* may be different.
2) **Dependence on different time period subsets**: It may be possible that the outcome (classification/forecasting) may not be captured by overall history representation. In most likelihood, it would be a subset of overall time-period representation that may impact the outcome.

A survey on sensor time series [138] mentions that the strategies used for time-series DA bear much resemblance to non-time-series DA, with two specific strategies for time-series DA – *input space* adaptation and *output space* adaptation.

#### 1) INPUT SPACE ADAPTATION

In the *input space* DA strategy, the impetus is to use/generate the source domain samples which resemble the target domain samples, much like reconstruction-based methods. Typically, prior knowledge (Wang et al. [139]) or GANs (Contra-GAN [140]) are used in this strategy.

#### 2) OUTPUT SPACE ADAPTATION

*Output space* DA strategy is used both for classification and forecasting (DAF [137]). In the case of classification Yang et al. [141], high-confidence labels on the target domain are selected for training, analogous to pseudo-semi-supervised training (refer to section Pseudo-Semi-Supervised Domain Adaptation). In the case of forecasting, domain-specific features are used (values of transformer network in DAF [137])

### J. EMERGING DOMAIN ADAPTATION FOR PRACTICAL SETTINGS AND REAL-WORLD CHALLENGES

Some models and techniques available in the literature do not fit into existing categories, have gained a lot of traction, and are, to some extent, very innovative and adapted to more practical settings and/or real-world challenges. These emerging DA techniques are mentioned below.

#### 1) FEW-SHOT DOMAIN ADAPTATION

The challenge with few-shot DA is that there is not enough target data that can conclusively conform to the simultaneous requirements of DA. These requirements are domain confusion and representation alignment between the two domains. One of the first works on few-shot DA is Motiian et al. [142]. They introduced Few-Shot Adversarial Domain Adaptation (FADA) using adversarial learning focusing on speed of adaptation. They alleviated the difficulty mentioned before by mixing source and target samples into four categories based on domain and class labels, and the classifier then worked on these four categories instead of the standard two. Further, they initialized the network (feature extractor and label classifier) using source data only, then updated the domain class discriminator (freezing feature extractor). Finally, they froze the domain class discriminator and updated the feature extractor and label classifier.

In the Domain-Adaptive Few-Shot Learning (DA-FSL) [143], they look to solve even a more complex problem related to few-shot learning, i.e., target data may have classes that can come from different domain. The focus of the domain-adversarial prototypical network (DAPN) in DA-FSL is to attain alignment in global domain distribution while keeping class discriminative-ness intact by introducing new losses (domain discrimination, domain confusion, classification). The losses are weighted using an adaptive

re-weighting mechanism. Another novel aspect was the use of attention before the embedding of the source.

Further, Yue et al. [144] proposed an end-to-end Few Shot Domain Adaptation method, which includes self-learning (called Prototypical Cross-domain Self-Supervised Learning (PCS) framework) and is unsupervised. The main idea is knowledge transfer from source to target is to find similarities between instance and prototype (representative), making the transfer more robust.

#### 2) ZERO-SHOT DOMAIN ADAPTATION

Zero-Shot DA is a complex scenario because actual target domain data is not present during training time; only some information about it (typically target metadata) is available. Zero-Shot DA differs from DG because DG does not have any information about the target data, not even the metadata.

1) **Zero-Shot Learning (usage of task-irrelevant data):** For the computer vision task, Peng et al. [145] used information in task-irrelevant data (domain pairs) to help understand network information about the non-available task-relevant target domain.
2) **Zero-Shot Learning (new labels in the target domain):** The intention is to learn ''different'' class labels in the target domain, given labels in the source. This is genuinely not a DA scenario, as the label domain is different in both source and target. An example mentioned in Kodirov et al. [146] is that the label ''Polar Bear'' can be represented as embedding vectors of 'has fur,' 'is white,' and 'eats fish.' Any semantic embedding that is close to these embedding vectors can help label effectively.

#### 3) LABEL SET DIFFERENCE IN DOMAINS

This perspective helps to close the category (label) gap in DA – it may be possible that the target label set may contain more (or open-set) or less (or partial) than the source. The typical DA scenario is called closed-set DA, where the label set in source and target is the same. The solution that supports both open-set and partial is called universal domain adaptation.

1) **Open (set) Domain Adaptation:** The Open DA idea by Saito et al. [147] uses an adversarial generative model where the generator creates samples different from the data boundaries of source samples. The feature extractor component can either align the features of the target domain within the boundaries of the source domain or push away from the boundaries; the samples pushed away from boundaries represent the unknown class. The separate-to-adapt strategy ([148]) progressively (coarse boundaries to finer boundaries) separates known classes and unknown classes and uses the adversarial discriminative method. Saito and collaborators again discuss open-set domain adaptation with a benchmark towards open-set classification in syn2real [149]. Pan et al. introduced Self-Ensembling with Category-agnostic Clusters (SE-CC) [150]), which helps in domain adaptation by looking at cluster distribution of unknown (new) classes, giving more understanding to the network to segregate between known and unknown classes and within known classes.
2) **Partial Domain Adaptation:** The source domain having a greater number of label classes than the target domain, i.e., Partial DA setting, leads to a problem of negative transfer. Partial Adversarial Domain Adaptation (PADA) [151] implements an adversarial discriminative method and aligns the feature distribution of two domains in a shared space. Further, it weighs down the importance of the extra class(es) of the source domain. Cao et al. [152] extend their previous work using Example Transfer Network (ETN), where the strategy of weighting down the class importance is different. It evaluates transferability and only transfers examples like the target domain.
3) **Universal Domain Adaptation:** Universal DA is one of the most complex DA scenarios to deal with, and the research attempts are very recent. The idea by You et al. [153] is typically to appreciate two elements – domain similarity (which helps to understand if the task can be supported) and prediction uncertainty. Domain similarity deduces samples coming from similar labels, while prediction uncertainty deduces the unknown class. It further includes aspects of partial domain adaptation strategies by the same research group and supports all settings – closed/partial/open-set variations. The training tries to find an optimum probability (that the sample is part of the source class) which can help segregate if data can be worked on; else, mark it as unknown. V. N. and Kundu et al. [154] support Universal DA by using a proxy of unobserved class (a hypothetical negative class) and therefore helps in class separability.

#### 4) CONTINUOUS / SEQUENTIAL / INCREMENTAL DOMAIN ADAPTATION

In a representative DA setting, the source data and target data are available during the training time. However, in real-world settings, target data may be made available as we progress on DA testing over time, or the target domain itself may change. In these settings, continuous (or sequential or incremental) DA is imperative.

1) **Online domain adaptation:** In the work of J. P. and Mancini [155], continuous domain adaptation is done using batch normalization for unsupervised domain adaptation. Sharing of network parameters happens between source and target (online) except for the batch normalization params. Batch normalization parameters are updated on the go (over time). This online DA strategy was used in robotics use where the objects were lit differently in different settings.
2) **Predictive and Online domain adaptation:** For unsupervised learning scenarios, Mancini et al. in

AdaGraph [156] focused on a predictive domain adaptation scenario with an online learning component. The system learned generalizing from annotated source images alongside unlabeled samples (with associated metadata) from secondary domains. AdaGraph is used to understand the domain-specific parameters, and it provides those parameters to batch normalization layers as part of predictive DA.

3) **Continuously Changing Domains**: Sometimes, the task involved is such that domains vary continuously (e.g., self-driving car driving on a sunny day, and suddenly it rains); we cannot treat the shift as discrete or static domains. Continuous Unsupervised Adaptation or CUA [157] learns to adapt to new distribution without not deviating (replay) from how it performed in previous distributions. CUA has an element of adaptation (Adapt Module) and memory (to replay if the same domain is countered again, called Replay Module).
4) **Continuously Indexed Domain Adaptation:** One of the drawbacks of the existing DA techniques is that they look to transfer knowledge between categorical (A and B) domains. However, in the real world, continuously indexed domains are involved in many tasks. Continuously Indexed Domain Adaptation or CIDA [158] conditions domain index distribution on a discriminator that models the encoding. Another variant of CIDA is Probabilistic CIDA (PCIDA); here, instead of the predicted domain index as output, it provides mean and variance for the domain.

#### 5) OPEN COMPOUND DOMAIN ADAPTATION (OCDA)

At times, there do not exist any clear boundaries amongst the source and multiple target domains. X. S. and Liu [159] concentrated on open compound domain adaptation (OCDA), where the target domain is a composite of numerous unlabeled and homogeneous domains. To bootstrap generalization, they used curriculum domain adaptation in a data-driven self-organizing fashion – understand easy-to-hard, based on domain gaps. OCDA also separates characteristics discriminative between classes from those specific to domains. The curriculum of domain-robust learning is constructed from the teased-out domain feature. Further, the use of memory modules increases the support for new domains. The knowledge transfer happens from the source domain to target domain instances, and also, the network can dynamically balance the memory-transferred knowledge and the input information. If the new domain is close to any source domain, it can work as a typical domain adaptation; in case of a difference, the memory module helps.

#### 6) SOURCE DATA RESTRICTIONS

There are conditions where data privacy is a concern or source data is not available. DA model that relies less on no source data (post model creation) is a boon in those conditions. For example, Source Hypothesis Transfer (SHOT) by Liang et al. [184] only uses the source model instead of the source data. The model aligns the source model with target data by learning target-specific features (uses information maximization and self-supervised pseudo-labeling).

Universal Source Free domain adaptation [154] and Federated domain adaptation [160] also aim to support DA where the availability of source data during training is unsure. V. N. and Kundu et al. [154] support Universal DA (closed, open set, and partial domain adaptation) and use synthetically generated hypothetical negative classes, which can act as a proxy for the unobserved class, knowledge of class separability, and category gap. In federated domain adaptation [160], model parameters are trained for each source note separately, converging at different speeds. The use of dynamic attention help understands the weightage of each source model. Federated domain adaptation also uses concepts of Domain Alignment, Domain Disentanglement, and Mutual information minimization.

#### 7) SELF-SUPERVISED LEARNING IN DOMAIN ADAPTATION

Self-supervised learning (including domain adaptation) is typically a two-step sequential process; the first process step includes unsupervised learning from a pretext task (in CV: rotation, image reorganization, implanting, colorization, etc.), which is used to understand intrinsic domain information (in CV: say semantic information of images in a particular domain). In the second process step, this learning is applied to a new task which further broadens it. Bucci et al. [161] implemented a similar process for object recognition across domains. The first task broadens the previous supervised learning of semantic labels, and the second task focuses on understanding the structure of the objects and their orientation. Given that label bias does not affect self-supervised learning, it can be used in partial (Bucci et al. [162]) and open-set (Bucci et al. [163]) DA areas.

#### 8) META-LEARNING IN DOMAIN ADAPTATION

Meta-learning (or learning-to-learn) represents algorithms that learn from the output of other algorithms. These sit one level above (can be visualized as outer loop algorithms) over the standard task algorithms and are vital in model selection and tuning processes. Li and Hospedales [164] implemented meta-learning for semi-supervised DA and multi-source DA; they also mentioned that meta-learning could be used for good initialization. Meta-learning in DA helps to increase evaluation metrics (positive impact) by 0.7% (DANN) to 2.5% (MCD). Another example in the speech domain is the adaptation of generative-based dialogue systems for unseen domains - Ribeiro et al. [165] improved DiKTNet (a dialogue model) adaptation to unseen domains using meta-learning. Meta-learning also finds use in domain generalization ([166], [167])

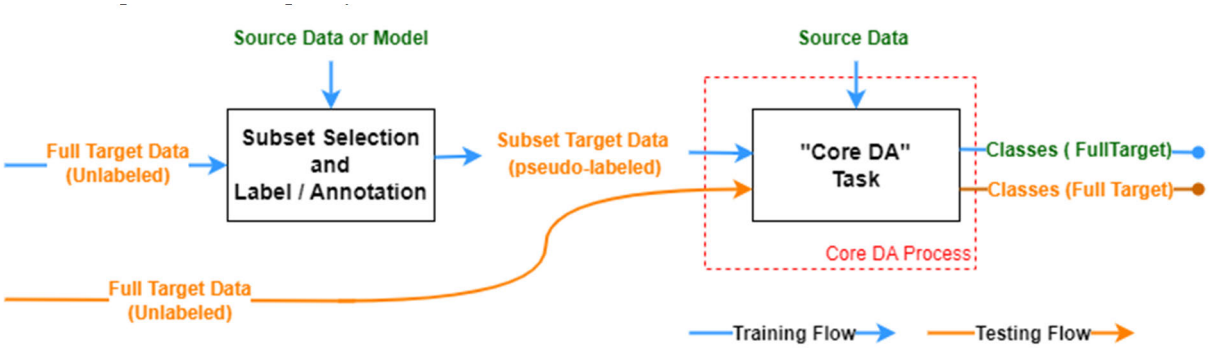


**FIGURE 10.** Typical Pseudo-semi-supervised DA strategy. A subset of target data is pseudo-labeled using "non-adapted" source model. The "core-DA" task ensures that the pseudo-labels assigned are corrected. Best viewed in color.

#### 9) PSEUDO-SEMI-SUPERVISED DOMAIN ADAPTATION

This set of methods includes the treatment of a subset of unlabeled target domain data and labeling them before the start of the "core" DA process (refer to Figure 10). Therefore, for the "core" DA process, there exists a subset of target domain data that is labeled and hence the name pseudo-semi-supervised DA. It may be noted that the initial labeling of unlabeled target domain data may be accurate or inaccurate, which is further refined during the "core" DA process.

1) **Active Learning in Domain Adaptation (Active DA):** While DA attains excellent results, the performances of DA methods often fall far behind their supervised counterparts. In such cases, active domain adaptation (Active DA) has recently gained a lot of interest. In the Active DA method, a subset of target samples is used to obtain annotations and further helps to improve the performance of the "core" DA. The focus is on selecting samples that not only include the diversity of target data but also represent the complexity.
   Su et al. [168], in Active Adversarial Domain Adaptation (AADA), used selection criteria based on diversity cue (dependent on optimal discriminator in adversarial setting) and uncertainty cue (dependent on cross-entropy, a proxy for empirical risk). They showed superior performance for digit recognition and object detection tasks. Prabhu et al. [169] further improved on basic active learning techniques of diversity cue and uncertainty cue by proposing Clustering Uncertainty-weighted Embeddings (CLUE). They weighted samples and selected them; here, diversity was supported by clustering and uncertainty by entropy weighting. They surpassed previous active learning-based SOTA (i.e., AADA) results in digit recognition and object detection.
2) **Pseudo-Labeling in Domain Adaptation:** Unlike active learning, Pseudo-label DA includes applying the model trained on labeled source data on a batch of unlabeled target data to predict labels / annotate. Here the labels/annotations on target data are not accurate but a reflection of labeled source data. Thereafter, one of the techniques is to train a new model with labeled source data and pseudo-labeled target data. However, this method has the inherent weakness of propagating noisy labels (incorrect labels).
   In CV, Kim and Kim [170], worked on abating the noisy label problem by implementing a joint optimization framework, i.e., iteratively updating the model (network) and pseudo-labels.
   In NLP, Wang et al. [171] used Generative Pseudo Labeling (GPL) for query-passage extraction purposes: where they retrieved positive passages from labeled data and applied that model for retrieving negative passages in target data. Thereafter, they used Margin-MSE loss which helped the cross-encoder to soft-label query-passage pairs effectively. They then used the soft-labeled pairs for the core task.
   In time-series, as part of the *output space* strategy, Yang et al. [141] selected high-confidence labels on the target domain for training.
   Moving Semantic Transfer Network (MSTN) [174] looked to align the centroid of each class in both labeled source and pseudo-labeled target data. Chen et al. [175], in Progressive Feature Alignment Network (PFAN), formulated an easy-to-hard strategy (ETHS) and used only an easy sample for downstream network (Adaptive Prototype Alignment or APA) use. ETHS and APA were then used iteratively till convergence for best results.

## V. DATASETS USED IN DOMAIN ADAPTATION

This section captures the existing and emerging datasets used for DA across CV, NLP, speech, time-series, and multi-modal data domains. One observation is that researchers use very few benchmark DA datasets, and the research is done in a very narrow set of tasks.

### A. COMPUTER VISION (CV) DATASETS

In Computer Vision (CV), most of the DA work has been done in digit recognition and image classification. Complex CV

**TABLE 11. Common computer vision (CV) datasets used in DA.**

| Tasks | Dataset | Dataset Description | Example DA References |
|---|---|---|---|
| Digit Recognition | MNIST [167] | Modified National Institute of Standards and Technology (MNIST) consists of 70K images (60K train+10K test) in 10 categories (digits) (28x28), monochrome. The digits have been centered and size-normalized in a fixed-size image. Created from the NIST database. | • DANN [49]<br>• DSN [31]<br>• DFA-MCD [168]<br>• Mean Teacher [82] |
| | MNIST – M [49] | Color photos patches of BSDS500 randomly extracted are combined as background to MNIST digits to form MNIST-M (Modified). It contains 59K+1 training and 90K+1 test image of 32x32. | • DANN [49]<br>• DSN [31]<br>• DRANet [176]<br>• TPN [47] |
| | U.S. Postal Service (USPS) [177] | Normalized (between -1 and 1) digit dataset from United States Postal Services handwritten scanned and segmented zip codes. 9229 samples in a total of 16x16. | • ADDA [48]<br>• DFA-MCD [173]<br>• MCD [178]<br>• TPN [47] |
| | SVHN [179] | Street View House Numbers, 32x32 RGB cropped, printed, normalized, and centered house number digits extracted from Google Street View. 630K and 530K (less challenging) samples. | • ADDA [48]<br>• Mean Teacher [84]<br>• CyCADA [64]<br>• TPN [47]<br>• DWT [111] |
| | Synthetic Digits / Numbers [49] | Synthetic numbers are created from Microsoft font. 32x32 images consist of various orientations, background colors, positions, blur, etc. Around 500K samples are present. | • DANN [49] |
| Image classification / Object Recognition | Office-31 [180] | 4110 images across 31 classes from 3 domains – Amazon (2817 total, 90 per class), DSLR (498 high-resolution images, 5 per class, different viewpoints (avg 3)), and Webcam (795 low resolution) | • FixBi [181]<br>• Contrastive Adaptation N/W [38]<br>• Generate to Adapt [182] |
| | Office-Home [183] | 15500 images across 65 classes and 4 domains – Art (artistic images/sketches), Clipart, Product (without background), Real-world. | • FixBi [181]<br>• SHOT [184]<br>• SPL [185]<br>• DWT [111] |
| Image classification / Object Recognition | Office – Caltech [186] | 10 categories (overlapping with the Office and Caltech256 dataset) with 2533 images. Vector quantized to 800 dimensions, and SURF BoW histogram features are also available for this dataset. | • DAN [39]<br>• CORAL [36] |
| Image Classification | DomainNet [72] | 569010 images across 6 Domains (clipart, real-world, sketch, infographic, painting, and quickdraw) include 345 categories (classes) of objects such as bracelets, planes, birds, and cellos. | • UAN [153]<br>• KD3A [187] |
| | ImageCLEF-DA [40] | ImageCLEF-DA consists of 50 images per 12 categories for each domain. This dataset was included in ImageCLEF 2014 domain adaptation challenge and includes 3 domains: Caltech-256 I, Pascal VOC 2012 (P), and ImageNet ILSVRC 2012 (I). | • DADA [60]<br>• SPL [185] |
| (Cross-Domain) Object Detection/ Classification | Syn2Real [149] | Syn2Real includes 3 datasets: *source* – synthetic renderings of 3D models, *validation* – images cropped from the Microsoft COCO dataset [188] and *testing/target* – images cropped from the Youtube Bounding Box dataset [189]. This is used for 3 tasks of closed-set classification (12 categories), open-set classification (12 and other categories) and detection. | • Syn2Real [149]<br>• Discriminative adversarial domain adaptation [190]<br>• Mean Teacher with Object Relations (MTOR) [191] |
| Image Classification/ Segmentation | VisDA-2017 [192] | Dataset for VisDA-2017 challenge into task areas – Image classification and segmentation. 280157 images (synthetic data to real imagery) across 12 categories. | • Contrastive Adaptation Network [38]<br>• JAN [40]<br>• TPN [47] |

**TABLE 11.** *(Continued.)* **Common computer vision (CV) datasets used in DA.**

| Tasks | Dataset | Dataset Description | Example DA References |
|---|---|---|---|
| Segmentation/ Object Detection | Cityscapes [193] | 5000 images (and 20000 additional images with coarse annotations) across 30 classes from real urban scenes for semantic segmentation from 50 different cities. A subset is used for object detection. | • CycleGAN [63]<br>• CoGAN [194]<br>• SAD [195]<br>• MTOR [191] |
| Object Detection | Foggy Cityscapes [196] | Simulated the dataset by adding fog to real scenes. It contains 20550 images based on Cityscapes [193] | • SAD [195]<br>• MTOR [191] |
| Segmentation | GTSRB [197] | German Traffic Sign Recognition Benchmark (GTSRB) contains images of 43 classes of traffic signs (rich background and varying light conditions), which are split into 39209 training images and 12630 test images. | • Mean Teacher [84]<br>• DAN [39] |
| | GTA5 [198] | 24966 synthetic images from Grand Theft Auto – V, across 19 classes (Cityscapes compatible) from a car perspective. | • ProDA [199] |
| | SYNTHIA [200] | SYNTHetic Collection of Imagery and Annotations has 9400 virtual city images across 13 classes | • Self-supervised Augmentation Consistency [201] |
| | Cityscapes Dataset [193] | Cityscapes have got semantic & original images (5000 High-Quality annotations, 20000 Coarse annotations) from 50 different cities. It serves as one of the benchmark datasets in road semantic/urban scenery segmentation | • Unsupervised Domain Adaptation method [202] |
| | Cross-City Dataset [202] | Cross-City Dataset has 4 cities with 100 labeled images and 1600 unlabeled images for each city in 647x1280 resolution | • Unsupervised Domain Adaptation method [202] |
| Video / Action Classification | HMDB51 [203] | HMDB51 contains 51 action categories (“kiss,” “jump,” “laugh,” etc., each having 101 clips) and a total of 6849 realistic videos from various sources, including web videos and movies. | • TAN3N [204]<br>• JAN [40] |
| | UCF101 [205] | 13320 videos (27 hours of clips) from 101 action categories. Realistic user-uploaded videos with a cluttered background. | • TAN3N [204]<br>• JAN [40] |
| Pose Estimation | LINEMOD [206] | An RGB+D dataset containing 1100, 15 object sequences with one object annotated (6D pose, class label, bounding box) | • DSN [31] |
| Gaze Estimation | MPIIGaze [207] | MPIIGaze consists of 213659 real images from 15 contributors | • DAGEN [208] |
| | EYEDIAP [209] | EYEDIAP provides images in RGB and RGB-D format. 16 contributors over 94 sessions provided images in 4 variations | • Domain Adaptation Gaze Estimation Network (DAGEN) [208] |
| 3D point cloud segmentation | SemanticUSL [210] | SemanticUSL contains 16578 unlabeled scans (training data) and 1200 labeled scans (evaluation data) | • LiDARNet [210] |
| | SemanticKITTI [211] | SemanticKITTI contains 23201-point clouds (training data) and 20351-point clouds (evaluation data) in 22 sequences | • LiDARNet [210] |
| | SemanticPOSS [212] | SemanticPOSS contains 2988-point clouds | • LiDARNet [210] |
| | LIBRE [213] | Real-world data features 3 environments & 10 LiDAR sensors – static targets, adverse weather & dynamic traffic. | • Survey on deep DA for lidar perception [214] |
| | DENSE [215] | High-quality, real-world camera+LiDAR dataset having 1150 scenes each 20 seconds. http://www.waymo.com/open | • Pseudo-labeling for Scalable 3D Object Detection [216] |

**TABLE 12. Common natural language processing (NLP) datasets used in DA.**

| Tasks | Dataset | Dataset Description | Example DA References |
|---|---|---|---|
| Sentiment Analysis (SA) | Amazon Reviews or Amazon Product Data [208] | A subset of 142.8 million Amazon product reviews of 4 product domains: Books (labeled -2000, unlabeled 6000), DVDs (labeled -2000, unlabeled 34741), Electronic items (labeled -2000, unlabeled 13153), and Kitchen appliances (labeled -2000, unlabeled 16785) is used. Adaptation amongst the twelve combinations ($\frac{4}{2}C$) experiments widely in research. Labeled: 1000 positive, 1000 negative for each category | • Stacked Denoising Autoencoder (SDA) [34]<br>• DANN [47]<br>• Neural Structural Correspondence learning (SCL) [64] [65] [66]<br>• Adversarial Memory Network (AMN) (Attention + DANN + SCL MemNet) [53]<br>• Asymmetric tri-training [85]<br>• Wasserstein Distance Guided Representation Learning (DANN) [52]<br>• SSL, Multitask tri-training [127]<br>• Domain and Task Adaptive pre-training (incl. multi-phase) [121] |
| | Airline Review [209] | 41396 reviews scrapped from Skytrax (www.airlinequality.com) in 4 categories: Airline, Airport, Seat, and Lounge. | • Neural Structural Correspondence learning (SCL) [65] |
| | REALNEWS [219] | Out of 11.90M articles from REALNEWS, a large corpus of Common Crawl's news articles 115000 (4 classes) are used. | • Domain and Task Adaptive pre-training (incl. multi-phase) [124] |
| | IMDb Reviews [220] | 50000 binary reviews from Internet Movie Database (IMDb). Polarized reviews, either review score < 4 or >7, is considered for the purpose. At most, 30 reviews per movie. | • Domain and Task Adaptive pre-training (incl. multi-phase) [124] |
| Text Classification (TC) | Fake News Challenge (www.fakenewschallenge.org) | 4 stances from the Fake news challenge: (unrelated), (discuss), (agree), and (disagree) were used to classify text (stance detection) | • A Variant of DANN [53] |
| | Fact Extraction and VERification (FEVER) dataset [221] | Based on Wikipedia classification, Fever Dataset has 55% (supported), 21% (refuted), 24% (Not Enough Information (NEI)) in total 185,445 claims. | • A variant of DANN [53] |
| Text Classification (TC) | Corpus of Historical American English (COHA) [222] | 450 million words from 1990 till date, COHA is the largest structured corpus of historical English. | • Adaptive Ensembling (Temporal) [86] |
| Text Classification (TC), Relationship Extraction (RE) | The New York Times Annotated Corpus [223] | Of 1.8 million articles in NYT Annotated Corpus, 4,800 articles with US POLITICS & GOVERNMENT descriptors were used. | • TC: Adaptive Ensembling (Temporal) [86]<br>• RE: Feature Extractor (piecewise CNN, GRU-based RCNN ), Attention, and Adversarial Training [224] |
| Part of Speech (POS) Tagging | Wall Street Journal portion of Penn Tree Bank (PTB-WSJ) [225] | Each sentence in the corpus is annotated with the Part of Speech (POS) tag. 45 Different POS Tag are used: sections 0-18 for training (38219 sentences, 912344 tokens), 19-21 for development (5527 sentences, 131768 tokens), and 22-24 for testing (5462 sentences, 129654 tokens) | • Adversarial Training – AT (DANN) [57]<br>• SSL, Multitask tri-training [130]<br>• AdaptaBERT [119] |
| | Treebank from Universal Dependencies (UD) v1.2 [226] | Multilingual treebank annotation in 33 languages (v1.2) for cross-lingual and NLP learning tasks. | • Adversarial Training – AT (DANN) [57] |
| | SANCL 2012 Shared Task Dataset [227] | Shared task dataset from parsing the web text from Google Web Treebank | • SSL, Multitask tri-training [130] |

**TABLE 12.** *(Continued.)* **Common natural language processing (NLP) datasets used in DA.**

| Tasks | Dataset | Dataset Description | Example DA References |
|---|---|---|---|
| | The Tycho Brahe Corpus of Historical Portuguese [228] | Tycho Brahe Corpus (TBC) contains text in the Portuguese language from 1502 to 1836, containing 1.4 million tokens. Different domains are created based on year bins. | • Marginalized structured dropout [229] |
| | PubMed-POS [16] | PubMed-POS includes 2 domains – PubMed (target typically) and WSJ portion of Penn tree-Bank (source typically) | • Frustratingly easy domain adaptation [18] |
| | PPCEME [230] | Penn Parsed Corpus of Early Modern English (PPCEME) is part of Penn Parsed Corpora of Historical English (PPCHE) and includes text samples and running text between early middle English and World War 1. Typically, Part of Speech annotated part of the dataset is used. | • AdaptaBERT [119] |
| Named Entity Recognition (NER) | CoNLL2003 [231] | Conference on Computational Natural Language Learning (CoNLL) 2003 data contains NER data in English and German and is the seminal NER dataset. English data is taken from Reuters corpus (1996-1997) and contains 1393 sentences and 4 categories (person, location, organization, and misc.) and 35089 named entities in CoNLL format. | • AdaptaBERT [119]<br>• Cross-Domain NER [232]<br>• Adversarially Trained Language Models [233] |
| | WNUT [234] | 2016 Workshop on Noisy User Text (WNUT) Named Entity Recognition in Twitter Shared task, has data in CoNLL format. The data is annotated with 10 fine-grained NER categories (person, geo-location, company, facility, product, music artist, movie, sports team, tv show, and other) | • AdaptaBERT [119]<br>• Adversarially Trained Language Models [233] |
| | LitBank [235] | 210,532 tokens that are evenly drawn from 100 English literary texts with entity and event annotations. | • Adversarial domain adaptation (ADA) (DANN + Contextual Embeddings) [56] |
| | TimeBank [236] | 183 English news articles containing annotations for events and temporal relations between them. | • Adversarial domain adaptation (ADA) (DANN + Contextual Embeddings) [56] |
| | BioNLP13PC [237] | BioNLP13PC contains 4 Named entities categories with 3 overlappings (Gene/ Protein, Chemical, Cellular component) with BioNLP13CG. It contains 5100 sentences and 15900 named entities in total. | • Cross-Domain NER [232] |
| | BioNLP13CG [237] | BioNLP13CG contains 16 Named entities categories with 3 overlappings (Gene/ Protein, Chemical, Cellular component) with BioNLP13PC. It contains 5900 sentences and 21300 named entities in total. | • Cross-Domain NER [232] |
| Named Entity Recognition (NER) | CBS SciTech News Dataset [232] | CBS SciTech News Dataset contains 620 articles from CBS SciTech News (https://www.cbsnews.com/) in 4 categories (person, location, organization, misc.) overlapping with CoNLL2003 entities with 4138 named entities. | • Cross Domain NER [232] |
| | FIN [238] | FIN dataset contains data from the Finance domain from public U.S. Security and Exchange Commission (SEC) filings. It includes 8 documents annotated with 4 categories of CoNLL2003 (person, location, organization, misc.) entities. | • Adversarially Trained Language Models [233] |
| | JNLPBA [239] | JNLPBA is a Biomedical domain dataset and is a subset of the GENIA corpus. It has around 22400 sentences and 5 entity types (protein, DNA, RNA, cell_type, and cell_line) | • Adversarially Trained Language Models [233]<br>• GreenBioBERT [240] |
| | BC2GM [241] | BioCreative II Gene Mention Recognition (BC2GM) Dataset contains 20100 sentences and is annotated with gene mentions. | • Adversarially Trained Language Models [233]<br>• GreenBioBERT [240] |
| | BC5CDR [242] | BioCreative V CDR corpus (BC5CDR) has 4409 annotated chemicals, 5818 diseases, and 3116 chemical-disease interactions and is derived from 1500 PubMed articles | • GreenBioBERT [240] |

**TABLE 12.** *(Continued.)* **Common natural language processing (NLP) datasets used in DA.**

| Tasks | Dataset | Dataset Description | Example DA References |
|---|---|---|---|
| | NCBI-disease [243] | • NCBI-disease dataset is created from 793 annotated PubMed abstracts by 14 annotators. It has 6892 disease mentions | • GreenBioBERT [240] |
| | BC4CHEMD [244] | • BioCreative IV Chemical compound and drug name recognition (BC4CHEMD) contains annotated 10000 PubMed abstracts with 84355 chemical entities. | • GreenBioBERT [240] |
| | LINNAEUS [245] | • LINNAEUS is a multiformat (XML, HTML, tab-separated) dataset containing documents from the biomedical domain primarily derived from the NCBI database, Medline, PMC, etc. | • GreenBioBERT [240] |
| Relation Extraction (RE) | English portion of ACE2005 dataset [246] | • The English portion of the ACE2005 dataset consists of 6 genres and 11 relation types.<br>• The genres are: Broadcast Conversation (bc), Broadcast News (bn), Newswire (nw), Telephone Speech (cts), Usenet Newsgroups (un), and Weblogs (wl) | • Genre Separation Networks (Domain Separation Networks) [62]<br>• CNN-based feature extractor followed by DANN [49] in [247] |
| | BioInfer [248] | • Bio Information Extraction Resource (BioInfer) contains 1100 sentences annotated with NER, syntactic dependencies, and relationships | • 2 Step method – Feature extractor (CNN or Bi-LSTM) frozen after the first step; GAN discriminator (Adversarial generative models) is used in the second step [249] |
| | AIMed [250] | • AIMed contains 750 real and 10000 synthetic MEDLINE extracts. Protein-Protein Interaction and Drug-Drug Interaction are present in these extracts. | • 2 Step method – Feature extractor (CNN or Bi-LSTM) frozen after the first step; GAN discriminator (Adversarial generative models) is used in the second step [249] |
| | DDI [251] | • Drug-Drug Interactions (DDI) is based on DrugExtraction Shared Task 2013 and contains Medline abstracts and DrugBank database Drug-Drug Interaction and documents, respectively. | • 2 Step method – Feature extractor (CNN or Bi-LSTM) frozen after the first step; GAN discriminator (Adversarial generative models) is used in the second step [249] |
| | UW Dataset [252] | • 20000 crowd-sourced instances based on Amazon Mechanical Turk and acquired using specific instructions and curation. | • Feature Extractor (piecewise CNN, GRU-based RCNN), Attention, & Adversarial Training [224] |

tasks (like pose estimation) are now getting traction. Table 11 lists common CV datasets used in DA in recent times.

### B. NATURAL LANGUAGE PROCESSING (NLP) DATASETS

Most of the domain adaptation work in NLP has happened for the sentiment analysis task. In recent years, more tasks have been explored. Table 7 lists common NLP datasets used in DA in recent times.

### C. SPEECH DATASETS

Table 8 provides a list of common speech datasets used in DA in recent times. We can see that most of the speech data domain DA work has happened in the speech recognition task.

### D. TIME-SERIES DATASETS

Table 14 mentions some time-series datasets used in DA. In industry, there is an expectation that many time-series-related DA problems would be there; however, the number of public time-series datasets used in DA continue to be very less.

### E. MULTI-MODAL DATASETS

The core field of multi-modal deep learning is developing, yet advances have been made in multi-modal DA. Table 15 provides common multi-modal datasets used in domain adaptation. The diversity of tasks is less, with the majority being Face Expression / Emotion recognition related.

## VI. CHALLENGES

Typical challenges of DA in the real-world and practical settings include:

1) **Few datasets in DA use**: Few datasets (Table 11 and Table 12) viz., MNIST, MNIST-M, SVHN, USPS, Office, and Amazon reviews) are typically used by researchers. There is a need to include more data sets – in the number of datasets and the size of data sets and develop a DA framework for specific applications. Further, the common datasets have fewer classes and

**TABLE 13. Speech datasets used in DA.**

| Tasks | Dataset | Dataset Description | Example DA References |
|---|---|---|---|
| Speech Recognition | Microsoft Cortana [253] | 3400 hours of close-talk and far-field speech | • Teacher-Student learning for unsupervised domain adaptation of Attention-based encoder-decoder [135] |
| | Aurora-4 corpus [254] | 7138 clean and noisy utterances together | • Sun *et al.* [255] |
| | Wall Street Journal (WSJ) [256] | WSJ0-+5000 word texts, WSJ1-20000 word texts, recorded in 2 settings for 80 hour | • Sun *et al.* [255] |
| | Librispeech [257] | 1000 hours from widely available audiobooks | • Sun *et al.* [255] |
| | SPEECON [258] | Multilingual data used for speech recognition in consumer devices. Italian data consists of 550 utterances by adult speakers in 4 microphone settings | • Denisov, Vu, and Font [259] |
| Speaker Adaptation | (D)APRA Resource Management (RM1) database [260] https://catalog.ldc.upenn.edu/LDC93S3B | Speaker Dependent (SD) [ 7344 recorded sentences, 12 speakers, 2 dialect sentences] and Speaker Independent (SI) [3360 recorded sentences, 80 speakers, 2 dialect sentences] are part of (D)APRA Resource Management (RM1) database | • Maximum likelihood linear regression (MLLR) [261] |
| Speech Activity Detection (SAD) | Fearless steps (FS) Fearless Steps Challenge corpus [262] | Cumulative 19,000 h of conversational speech coming from the Apollo 11 mission | • Deep CORAL and pseudo-labeling techniques [131] |

**TABLE 14. Time-series datasets used in DA.**

| Tasks | Dataset | Dataset Description | Example DA References |
|---|---|---|---|
| Mortality Prediction | MIMIC-III (Adult-AHRF dataset) [263] | Critical care database containing 58000 admission records (38645 adults and 7875 neo-natal) | • Variational Recurrent Adversarial Deep Domain Adaptation (VRADA) [264] |
| | Child-AHRF dataset [265] | Child-AHRF contains a record of 398 children (admission records) | • Variational Recurrent Adversarial Deep Domain Adaptation (VRADA) [264] |
| Advanced Driver Assistance Systems (ADAS) | Brains4cars [266] | Brains4cars contains 1180 miles of freeway driving for 10 drivers | • Domain Adversarial Recurrent Neural Network (DA-RNN) [267] |
| Fault Detection | Boiler Fault Detection Dataset [268] | Data for 3 boilers for over 2 years (2014-2016) | • Sparse Associative Structure Alignment (SASA) [268] |
| Air Quality Forecast | Air quality forecast dataset [269] | The dataset consists of air quality data, meteorological data, and weather forecast data covering 4 Chinese cities with each hour data for the 2014-2015 years | • Sparse Associative Structure Alignment (SASA) [268] |

instances. Results shown by researchers on diverse datasets would promote the creation of more datasets, and finally, the diversity would lead to capturing more practical settings.

**TABLE 15. Common multi-modal datasets used in DA.**

| Tasks | Dataset | Dataset Description | Example DA References |
|---|---|---|---|
| Face Expression/ Emotion Recognition | IEMOCAP [270] | The Interactive Emotional Dyadic Motion Capture (IEMOCAP) has 5 male and 5 female actors in a dyadic setting and up to 12 hours of audio-video clips. The dataset has 8 categories of emotions – Anger, excitement, Fear, Frustration, Happiness, Neutral, Sad, and surprise – present. | • Multimodal domain adaptation neural networks (MDANN) [89] used this as a source dataset.<br>• HMTL (Heterogeneous Modality Transfer Learning) [100] |
| | AFEW [271] | Acted Facial Expressions in The Wild (AFEW) has 957 labeled videos of acted facial expressions. The dataset has 7 categories of emotions – (Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise) | • Multimodal domain adaptation neural networks (MDANN) [89] used this as the target dataset. |
| | MSP-IMPROV [272] | MSP-IMPROV has approximately 8500 audio-visual dyadic recordings in 4 categories Natural interaction, Other-improvised, Target-Improvised, Target-read with emotions of Angry, Happy, Neutral, and Sad. | • [95] used inter categories as the source and target domains |
| | BUAA-VISNIR [273] | BUAA across 2 modalities (VIS, NIR) has 150 subjects | • Missing Modality Transfer Learning via latent low-rank constraint (M2TL) [97] |
| | Oulu-VISNIR [274] | Oulu, across 2 modalities (VIS, NIR), has 80 subjects having 6 expressions (anger, disgust, fear, happiness, sadness, and surprise) | • M2TL [97] |
| | CMU-PIE [275] | CMU Pose Illumination and Expression (PIE) dataset consists of 41368 images of 68 people in 13 poses, 43 illuminations, and 4 expressions | • M2TL [97] – Pose C27 is used |
| | Yale B [276] | Yale B Face dataset consists of 2414 images of 38 people (about 64 images per person), different illumination and expressions | • M2TL [97] |
| | ALOI-100 [277] | Amsterdam Library of Object Images (ALOI-100) consists of over 100 images with a total of 110250 images of objects | • M2TL [97] |
| | COIL-100 [278] | COIL-100 (Columbia Object Image Libraries 100) consists of 72 images of 100 object classes, i.e., a total of 7200 images | • M2TL [97] |
| | CREMA-D [279] | CREMA-D consists of 7442 clips from 43 female and 48 male actors across various ethnicities. Each actor's data includes 12 sentences across 4 emotion levels (L, M, H, and unspecified) and 6 emotions of Anger, Disgust, Fear, Happy, Neutral, and Sadness. | • Domain Adaptation Conditional Semi-Supervised Generative Adversarial Networks (dacssGAN) [98] |
| | RAVDESS [280] | The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) consists of 7356 recordings of songs with emotions of calmness, happiness, sadness, anger, and fear and expressions of calm, happy, sad, anger, fear, surprise, and disgust from 24 actors | • Domain Adaptation Conditional Semi-Supervised Generative Adversarial Networks (dacssGAN) [98] |
| | CMU-MOSI [281] | CMU Multimodal Opinion Sentiment Intensity (CMU-MOSI) consists of opinion video clips (2199), with clips rated in sentiment from -3 to +3 | • HMTL (Heterogeneous Modality Transfer Learning) [100] |
| Image Captioning | NUS-WIDE [282] | NUS-WIDE 269648 images and 5018 tags were collected from Flickr. Objects and scenes are manually annotated with 81 concepts | • Deep Transfer Networks (DTN) [96] used 10 domains – birds, buildings, cars, cats, dogs, fish, flowers, horses, mountains, and planes. |
| Event Rumor Detection | PHEME [283] | Approximately 5800 threads of about 5 events are classified into Rumors and Non-rumors. Further, if rumor, then it is further classified into true, false, & unverified. | • Multimodal Disentangled Domain Adaption (MDDA) [93] |
| | PHEME_veracity [284] | Extends PHEME datasets by 4 more events. 6425 rumor threads with 9 events are classified into Rumors and Non-rumors. Further, if rumor, then it is further classified into true, false, and unverified. | • Multimodal Disentangled Domain Adaption (MDDA) [93] |
| Action Recognition | EPIC-Kitchens [285] | EPIC-Kitchens consists of 100 hours, about 20M frames, 700 variable-length videos of 90K actions in 45 environments | • Multi-Modal Self-Supervised Adversarial Domain Adaptation or MM-SADA [94] |

**TABLE 16.** Real-world challenges for DA.

| S. No. | Real-world challenge | Reference DA methods/techniques to cater to challenges | Relevant DA case-studies |
|---|---|---|---|
| 1 | **Availability of labels in the target domain**: As data availability becomes difficult, the availability of labeled data is even more difficult. | • Semi-supervised DA: Support of few labeled data in the target domain<br>• Unsupervised DA: Support for no-labeled data in the target domain | • Section VII extensively discusses case studies for semi-supervised and unsupervised DA |
| 2 | **Supporting Label Set Difference or Category Gap**: Ability to support the difference in Label set; target domains may have lesser labels (partial-set DA), more labels (open-set DA), or both (universal DA). Need to deal with two aspects together – the category gap and domain gap (shift) | • Partial-set DA: [151], [152], [286]<br>• Open-set DA: [147], [148]<br>• Universal DA: [154], [153] | • Partial-set DA: Predicting protein functions for new proteins (target function classes are limited) [152]<br>• Open-set DA: understanding unknown class (openness) in target domains in any setting [147]<br>• Finding animals in the wild which were not part of the dataset [153] |
| 3 | **Learning with a lesser target or no data:** Learning quickly with no data (zero-shot DA) or few (few-shot DA). This ask is different from semi-supervised or unsupervised DA in that semi, or unsupervised DA does see the data at training time, but it is not labeled; however, zero-shot and few-shot hardly see any data or some data, respectively. | • Zero-shot DA: [145]<br>• Few-shot DA: [142] [143]<br>• Domain Generalization ( [112], [287]) | • Zero-shot DA: Industrial application where given CAD models (texture less), train an RGD object classifier [145]<br>• Few-shot DA: Adapting to natural images (which are less in number) when synthetic images are available [143] |
| 4 | **Learning on the Go:** Learning and adapting to new data as data in the target or multiple domains are presented to the model. | • Online DA: [155]<br>• Continuously changing domains: [157] | • Online DA: Robot deployment, where a robot learns as it explores the environment [155]<br>• Continuously changing domains: Video game learning (reinforcement learning and classification), weather changes (from dry to drizzle to rainy to cloud burst) [157] |
| 5 | **Availability of source data during testing:** Due to various reasons like privacy, etc., source data is not available for training when target data is available. | • Source Free DA: Use of source model only**:** [184]**,** [154]<br>• Federated DA: [160] | • Source Free DA: Access to source data is restricted for privacy reasons (banks, hospitals) or is decentralized [184]<br>• Federated DA: Applied in areas where access to data is restricted, and there is also a domain shift in data. |
| 6 | **Metadata is available in the target domain but not actual data.** Actual data gathering may be costly, but metadata gathering may be less costly. | • Predictive Domain Adaptation: [156] | • Predictive Domain Adaptation: Ability to understand and adapt to portraits varying over the years and geographically with only some information related to years and geography available [156] |
| 7 | **Imbalanced Data**: Some real-world problems include data imbalance / long tail (e.g., over-representation of English in words, under-representation of females in images, etc.), where classes are not evenly distributed in the source and target domain. | • [84] in MNIST to SVHN adaptation used a class balance term for improving accuracy. However, this is not understood in detail as to how this benefits DA.<br>• [288] | • Imbalanced Dataset support: Support imbalanced data taken over time, different sensors or users [288] |
| 8 | **Support for Multiple Task:** Multi-Task DA is a scenario where the same dataset is used to perform multiple tasks simultaneously. Example: Semantic Segmentation and Object Detection at the same time. | • Multi-Task Domain Adaptation: [289] | • Multi-Task Domain Adaptation: Helps understand the hierarchy/grouping amongst the data elements (Same make of car / same body type). Honda make images may be available and Sedan images – one could understand Honda Sedan image [289] |
| 9 | **Out-of-domain testing to understand the brittleness of models/methods:** There is an inherent need to understand the robustness of models/methods by testing them on out-of-domain. Dedicated out-of-domain datasets are not present currently. | • In-domain and Out-of-domain results: [290] | • Long context Q&A can help increase the robustness of models as it typically has out-of-domain aspects included. [290] mentioned the chatbot, which should understand the context of n previous questions above in the Q&A (NLP) area. A similar aspect should be available to test DA, too, for the robustness of models. |
| 10 | **Sequential or Lifelong Learning:** [33] mentions Functional, Relational, and representational aspects of Lifelong learning. Memory-based approaches are integral to Lifelong learning | • Memory function in DA: [55] | • DA applications should be able to build on their previous memory and relate to the memory – for example: in quality control, the system should be able to remember a quality control issue aspect which is like the new issue and work accordingly |

**TABLE 16.** *(Continued.)* **Real-world challenges for DA.**

| S. No. | Real-world challenge | Reference DA methods/techniques to cater to challenges | Relevant DA case-studies |
|---|---|---|---|
| 11 | **Support for Multimodal data:** Multiple modalities, when combined or fused, should increase the performance of DA. Designing effective fusion strategies – an early fusion of features or late fusion for task decision – remains a challenge. | • Images with different modalities: [291]<br>• Multimodal DA neural networks (MDANN): [89]<br>• HMTL (Heterogeneous Modality Transfer Learning): [100] | • Most real-world data is multimodal; more effective strategies must be developed to increase the performance of DA. |
| 12 | **Self-Learning mechanisms:** To have more positive transfer, there is a need for a self-learning mechanism where models can learn better domain intrinsic information and further increase the accuracy. Also, it would aid the adoption of DA just like self-supervised models have been done in NLP | • Pretext task followed by DA: [161], [162] [163] | • In Partial and Open-set domain adaptation settings, where the number of labels/classes in target data is unsure. |
| 13 | **Building representative systems / Mitigating Bias in AI/ML systems:** Due to the kind of data used, there exists a bias in AI/ML systems which exaggerates stereotypes, the relationship between color and race | • Domain adaptation followed by Elastic Weight Consolidation / lattice-rescoring [292] | • Understand product diversity across geographical locations – soap in western countries is liquid, while it is solid in non-western countries [293] |

2) DA has the promise to apply to real-world problems and solve them. Researchers have started investigating and solving some of the challenges, and some are yet to be explored. Table 16 provides a view of real-world challenges and examples of research work undertaken; however, some areas are still to be examined.
3) **Need for more tasks and applications**: New/other applications involving different types of data (like NLP [115], [33]) for DA can be understood. Time-series data adaptation is not looked at much (sensor type adaption may be a great use case). Further, multi-modal data-related domain adaptations are few. Also, industrial applications (where the target is industrial data) can be looked at by exploiting domain adaptation (source data is academic data). There is a need to develop a DA framework in these areas.
4) **Research bias for Classification tasks:** In computer vision (refer to Table 11), most of the work done is in classification tasks (digit recognition and image classification). Other tasks (pose estimation, object detection, etc.) are less explored.
   Similarly, most of the work reported in NLP domain adaptation is in sentiment analysis, followed by classification tasks (as in CV) (refer to Table 12), and not many tasks (most are 1:1 adaptation tasks) are explored by researchers on the techniques published by them. Areas like dependency parsing (DEP), Named Entity Recognition (NER), part-of-speech (POS), and other areas are explored significantly less.
5) **Bidirectional DA:** It is understood that DA from the source domain to the target domain may yield good performance, but the reverse (i.e., target domain to source DA) may not yield that good performance. Few papers discuss the bidirectional results. Reasons are not understood as to why a particular direction yields better performance over the other direction. Example: SVHN to MNIST accuracy is very high [84], while MNIST to SVHN is not very high. A general-purpose strategy is required for bi-directional DA.
6) **Effective comparison metrics missing for some DA scenarios:** Typically, absolute mAP is used for object detection tasks – however, it is the relative mAP (source-only baseline and after DA) that is important for DA. It is much better than absolute mAP as different papers also use models trained with different hyperparameters. There is a need of similar effective comparison metrics.
7) **Varied model and data parameters in DA:** Fair and comprehensive evaluation of DA approach and reusability comparison is difficult due to varied metrics, hyper-parameters and data input (e.g., image size). There is an imperative need of standardization of some possible parameters e.g., image size.

## VII. APPLICATIONS OF DOMAIN ADAPTATION

Given that DA includes relevant elements and supports generalization, it has found usage in many applications. Mentioned are some motivating examples and possible usage in the future.

### A. COMPUTER VISION (CV) DOMAIN ADAPTATION USAGE

DA in CV continues to mirror the progress of CV tasks and techniques with a lag. The initial focus of DA in CV was on simple CV tasks – like digit recognition and image

classification, but later, the focus included complex tasks of object detection, segmentation, depth estimation and similar. Surveys have been done on domain adaptation on specific computer vision tasks, e.g., semantic segmentation [294] and object detection [295]. The current focus is increasingly on even more complex tasks (e.g., pose estimation, video classification), complex datasets (e.g., in the wild, 3D), improve state-of-the-art DA metrics in previously mentioned tasks. Also, due to the scarcity of data in the target domain, most DA methods adapt from synthetic or other domain data to real data.

Most of the work on DA in CV is on 2 Dimension (2D) data, e.g., camera images, followed by 2D data with time, e.g., video images, followed by a focus on 3D, e.g., LiDAR (Light Detection And Ranging). A survey on LiDAR perception by [214] further captures deep DA techniques.

Table 17 provides a view of different CV tasks and key DA advances in those specific tasks. These tasks and techniques have found much use of DA in the CV in industries (further discussed in the section Industrial Applications), e.g., AI imaging is widely used in the healthcare sector while LiDAR DA is used in Advanced Driver Assistance Systems (ADAS) or Autonomous driving. These techniques are also used in situations where the data is derived from different foundations (geographic, genetic, cultural, age, etc.)

### B. NATURAL LANGUAGE PROCESSING (NLP) DOMAIN ADAPTATION USAGE

Similar to CV, DA in NLP also mirrored NLP task and technique progress with a small lag. Recurrent Neural Network (RNN) based models (including LSTM) are of much use for NLP settings. Initial research in NLP focused on improving the embedding layer and vocabulary difference between source and target domain. Thereafter, adversarial-based methods (including GANs) were also employed for the NLP domain adaptation task. Post-2017, after the advent of Attention and Attention-based Transformers [5], considerable NLP research has been done as to how to use pre-trained models for the task at hand. This deviated from the typical DA technique where both source and target domain data were available at once; in the case of pre-trained models, source data was not available.

NLP and DA in NLP have been popular in the industry because data creation is much easier than CV – there is no need for a camera in business process; further, much of the data is generated in the form of social media content, literature by authors and as news articles. Tasks like sentiment analysis, text classification, natural language inference, language identification, part-of-speech (POS) tagging, dependency parsing, named entity recognition (NER), Question and Answers (Q&A), relation extraction (RE), neural machine translation (NMT), Sentence specificity prediction are used in document and information focused

**TABLE 17. DA usage in various computer vision (CV) areas.**

| CV Tasks | Example DA References |
|---|---|
| Digit Recognition | • DFA-MCD [173]<br>• DRANet [176]<br>• CLUE [169]<br>• Active Adversarial Domain Adaptation (AADA) [168]<br>• Moving Semantic Transfer Network (MSTN) [172]<br>• Mean Teacher [84]<br>• MCD [178]<br>• CyCADA [64]<br>• ADDA [48]<br>• DANN [49]<br>• DSN [31] |
| Image classification / Object Recognition | • FixBi [181]<br>• DAPN [143]<br>• SHOT [184]<br>• SPL [185]<br>• KD3A [187]<br>• DADA [60]<br>• UAN [153]<br>• Contrastive Adaptation Network [39]<br>• Moving Semantic Transfer Network (MSTN) [172]<br>• Generate to Adapt [182]<br>• Face Recognition in Unlabeled Videos [296]<br>• CORAL [37]<br>• DAN [40]<br>• Syn2Real [149]<br>• Mean Teacher with Object Relations (MTOR) [191] |
| Image Classification/ Segmentation | • Contrastive Adaptation Network [39]<br>• JAN [41] |
| Object Detection | • CLUE [169]<br>• Progressive Domain Adaptation [297]<br>• Active Adversarial Domain Adaptation (AADA) [168]<br>• Domain Adaptive Faster R-CNN [298]<br>• Deep Intelligent Visual Surveillance (DIVS) ( [299]<br>• Syn2Real [149] |
| Segmentation / Object Detection | • CycleGAN [63]<br>• CoGAN [194] |
| Segmentation | • [294]<br>• DA-DETR (Transformer-based) [300]<br>• ProDA [199]<br>• Self-supervised Augmentation Consistency [201]<br>• Domain transfer through deep activation matching [51]<br>• Class-based self-testing [88]<br>• Mean Teacher [84]<br>• GAN-based approach [301]<br>• Pixel-level Adversarial and Constraint-based Adaptation [50],<br>• DAN [40] |
| Image-to-Image Translation | • Different Adaptation Methods for Neural Style Transfer [302] |

**TABLE 17.** *(Continued.)* DA usage in various computer vision (CV) areas.

| CV Tasks | Example DA References |
|---|---|
| Video/ Action Classification | • TAN3N [204]<br>• JAN [41] |
| Depth Estimation | • AdaDepth [303] |
| Pose Estimation | • DSN [31] |
| Gaze Estimation | • Domain Adaptation Gaze Estimation Network (DAGEN) [208] |
| 3D point cloud segmentation | • LiDARNet [210]<br>• Survey on deep DA for lidar perception [214]<br>• Pseudo-labeling for Scalable 3D Object Detection [216] |
| Image Captioning (Also, cross-modal) | • Domain critic and multi-modal critic [304]<br>• Dual Learning [305]<br>• Densecap [306] |

industries where a lot of text data is generated due to the business processes involved. Tasks discussed in Table 18 are used in NLP applications widely in the industry (industrial applications are further discussed in the section Industrial Applications).

### C. SPEECH DOMAIN ADAPTATION USAGE

Most of the DA work in the speech area is in Automatic Speech Recognition (ASR). Environment noises are the main culprit that a model trained on the manually collected dataset (source) does not perform in real-world data (target) in ASR. Table 19 mentions many references as to how DA is used to address this mismatch and enhance quality. There is also work related to Text to speech (TTS) translation that employs DA for increasing application domain and robustness.

### D. TIME-SERIES DOMAIN ADAPTATION USAGE

The main idea of using DA in time-series data is to learn temporal latent representations of time-series data that are domain-invariant. However, learning the temporal representations is an arduous task due to dependency amongst timestamps, and a change in lags/offsets leads to difficulty in extracting domain-invariant representation. Table 20 provides a view of how DA is used to solve two major time series tasks of classification and forecasting.

DA is used to improve the performance of time series systems in healthcare [264], Driver assistance systems [267], and others [319]. Also seen is a movement from univariate time series to multivariate time series problem-solving.

### E. MULTI-MODAL DOMAIN ADAPTATION USAGE

Domain adapting multimodal data is very much relevant as much real-world data is multi-modal. Multi-modal DA

**TABLE 18.** DA usage in various natural language processing (NLP) areas.

| NLP Tasks | Example DA References |
|---|---|
| Sentiment Analysis (SA) | • Domain and Task Adaptive pre-training (incl. multi-phase) [121]<br>• Neural Structural Correspondence learning (SCL) [66] [67] [68]<br>• Wasserstein Distance Guided Representation Learning (DANN) [54]<br>• SSL, Multitask tri-training [130]<br>• Adversarial Memory Network (AMN) (Attention + DANN + SCL MemNet) [55]<br>• Asymmetric tri-training [87]<br>• DANN [49]<br>• Stacked Denoising Autoencoder (SDA) [34] |
| Text Classification (TC) (non – SA) | • A variant of DANN [53]<br>• Adaptive Ensembling (Temporal) [86]<br>• TC: Adaptive Ensembling (Temporal) [86] |
| Part of Speech (POS) Tagging | • AdaptaBERT ( [128]<br>• Adversarial Training – AT (DANN) [57]<br>• SSL, Multitask tri-training [130]<br>• Marginalized structured dropout [229] |
| Named Entity Recognition (NER) | • Adversarially Trained Language Models [233]<br>• Adversarial domain adaptation (ADA) (DANN + Contextual Embeddings) [56]<br>• Adversarially Trained Language Models [233]<br>• AdaptaBERT [128]<br>• Cross-Domain NER [232]<br>• GreenBioBERT [240]<br>• DSN [52] |
| Question and Answers (Q&A) | • Generative Pseudo Labeling [170]<br>• Semi-supervised QA with generative domain-adaptive nets [307] |
| Relation Extraction (RE) | • 2 Step method – Feature extractor (CNN or Bi-LSTM) frozen after the first step; GAN discriminator (Adversarial generative models) is used in the second step [249]<br>• Genre Separation Networks (Domain Separation Networks) [62]<br>• CNN-based feature extractor followed by DANN [49] in [247]<br>• Feature Extractor (piecewise CNN, GRU-based RCNN), Attention, and Adversarial Training [224] |
| Neural Machine Translation (NMT) | • Effective domain mixing for neural machine translation [308]<br>• Cost Weighting [309]<br>• A Survey of DA for NMT [310] |
| Sentence specificity prediction | • Domain agnostic real-valued specificity prediction [311] |
| Image Captioning (also, cross-modal) | • Domain critic and multi-modal critic [304]<br>• Dual Learning [305]<br>• Densecap [306] |

**TABLE 19.** DA usage in speech areas.

| Speech Tasks | Example DA References |
|---|---|
| Speaker Recognition | • Unsupervised Domain Adaptation via Domain Adversarial Training for Speaker Recognition [312] |
| Automatic Speech Recognition (ASR)/ End-to-End Speech Recognition | • DUST (self-training and pseudo-learning based) [313]<br>• GRL and DSN-based architectures [132]<br>• GAN-based [133]<br>• Teacher-Student learning for unsupervised domain adaptation of Attention-based encoder-decoder [135]<br>• Augmented Cyclic Adversarial Learning (ACAL) [314]<br>• Domain Adaptation of End-to-end Speech Recognition in Low-Resource Settings [315]<br>• Unsupervised domain adaptation by adversarial learning for robust speech recognition [259]<br>• An unsupervised deep domain adaptation approach for robust speech recognition [255]<br>• Knowledge Distillation based [134]<br>• Dauto [316] |
| Speech Activity Detection (SAD) | • Deep CORAL and pseudo-labeling techniques [131] |
| Text to Speech (also Multimodal) | • TDASS [317] |
| Speaker Adaptation | • Maximum likelihood linear regression (MLLR) [261] |
| Speech Enhancement | • Low-rank sparse decomposition model followed by DA [318] |

systems can support missing modalities in target data ([315], [100]), and the adaptation process is much more robust than unimodal DA, reinforcing that AI and ML systems can improve by learning from multiple

DA has been used in various multi-modal settings, i.e., tasks and modalities (refer to Table 21). The advances made here mirror the advances in individual modalities and other trends (e.g., knowledge distillation).

### F. INDUSTRIAL APPLICATIONS

Domain Adaptation has been widely adopted by the industry and is of relevance in Industry 4.0. Table 22 provides different use cases and how DA is used.

DA has uses in cross-industry and industry-specific use cases. DA drastically reduces not only the data requirements but also the number of machine learning / artificial intelligence models. This leads to reduced capital expenditure (CAPEX) and upfront effort. The reduced CAPEX is due to truncated activities of data procurement, data annotation, multiple model training etc. Further, a decrease in operational expenditure (OPEX) costs and efforts is guaranteed as Machine Learning Operations (MLOPs) efforts are reduced.

**TABLE 20.** DA usage in time-series tasks.

| Time-series Tasks | Example DA References |
|---|---|
| Classification | • Soft Parameter Sharing [320]<br>• Optimal Transport methods [321]<br>• Sparse Associative Structure Alignment (SASA) [268]<br>• Convolutional deep Domain Adaptation model for Time Series data (CoDATS) [322]<br>• Classification of Satellite Image Time Series [323]<br>• Domain-Adversarial Recurrent Neural Networks [267]<br>• Augmented Cyclic Adversarial Learning (ACAL) [314]<br>• Variational Recurrent Adversarial Deep Domain Adaptation (VRADA) [264]<br>• DAuto [316] |
| Forecasting | • Domain Adaptation Forecaster (DAF) [324]<br>• Sparse Associative Structure Alignment (SASA) [268] |

**TABLE 21.** DA usage in multi-modal settings.

| Tasks | Modalities | Example DA References |
|---|---|---|
| Event Classification | • Video<br>• Audio | • Audio-Visual Emotion Recognition [95]<br>• Audio-Visual Sentiment Analysis [100]<br>• Multimodal DA neural networks (MDANN) [89] |
| | • Text<br>• Image | • Domain Adaptation Forecaster (DAF) [324]<br>• Sparse Associative Structure Alignment (SASA) [268] |
| Text to Speech | • Text<br>• Speech (Audio) | • TDASS [317] |

MLOPs efforts that are reduced involve monitoring, retraining, versioning, and serving- all because of the lesser number of domain-adapted models.

## VIII. FUTURE RESEARCH FRONTIERS

The future research frontiers must look at solving the challenges mentioned in section VI. Also, the body of research in DA is currently focusing on

- **Including state-of-the-art (SOTA) techniques or methods**: Experiments and evolution of research in other areas of deep learning are flowing into DA. Not only are researchers looking to support DA in the base technique but also, they are looking to use a derivate of the technique to enhance DA. For example, attention and transformer-based models have found

**TABLE 22. Industrial applications: DA use-cases across industries.**

| S. No. | Industry | Area / Sub-Area | DA Use-case | DA Technique / Method |
|---|---|---|---|---|
| 1 | Across Industries | Customer Service | • Adapting to accent or speech: Commercial speech recognition software sometimes cannot understand the accents or speech of particular people. DA can help generalization across speakers.<br>• Adaptation of generative-based dialogue systems for unseen domains [165] | • Contrastive Learning with mutual Information Maximization (CLIM) on Airline's review dataset [218] in [325]<br>• [165] used meta-learning-based DA on MultiWOZ [326] dataset |
| 2 | Across Industries | Personalization | • Zhang *et al.* [327] and Yang *et al.* [328] used DA in personalizing responses by conversational robots in a two-step process. The first step is learning from a larger dataset, while the second step is fine-tuning based on small-scale personal conversation data. Similar aspects can be extended to other areas of personalization. | • Pre-training is followed by fine-tuning in Zhang *et al.* [327] and Yang *et al.* [328] |
| 3 | Across Industries | Price Forecasting | • Jin *et al.* [137] applied Domain Adaptation Forecaster (DAF) on time-series data to understand the price of a domain based on the price of another domain, where data is abundant | • DAF - Shared attention module and domain discriminator between two domains, with private encoders and decoders (Novel Design) |
| 4 | Across Industries | Quality Assurance or Inspection | • MMD DA technique is used in semiconductor manufacturing to adapt training and test data, which have deviations from the manufacturing process [329].<br>• Thota *et al.* [330] used multi-source DA to identify and verify the presence of use-by-date information in retail food packaging. | • Usage of Discrepancy-based Method DAN [40] (CNN followed by MMD) in [329]<br>• Usage of Discrepancy-based Method DAN [40] and Multisource domain adaptation in [330] |
| 5 | Across Industries | Robotics | • Many robotic applications, once they are deployed, need to adapt to various small shifts over time. Wulfmeier, Bewley, and Posner [331] provide as to how these small shifts (like changes in lighting and weather) can be supported by DA.<br>• Bousmalis *et al.* [332] used pixel-level domain adaptation (GraspGAN) and reduced the number of samples required to achieve a given performance level by 50 times. | • Usage of Adversarial models, like ADDA [48] in [331]<br>• Usage of Adversarial Models – used DANN [49] (feature level) and GraspGAN (pixel level) in [332] |
| 6 | Automotive | Autonomous Driving | • Adaptation to new seasons, roads, and geographic areas.<br>• Barbato *et al.* [333] compare the efficacy of DA to supervised training in semantic segmentation using a new measure; further, they enforce joint constraints between source and target features.<br>• Kothandaraman, Chandra, & Manocha [334] introduced BoMuDANet as an unsupervised adaptation method of visual scene understanding in different and unstructured driving conditions.<br>• Munir, Azam, & Jeon [335] proposed Self Supervised Thermal Network (SSTN), which understands the co-occurrence of information in multiple sensor data environments using a contrastive learning approach in one of the stages. | • Multiple constraints (loss) for domain invariant features and cross-entropy for class discriminative-ness in [333]<br>• Novel self-training "Alternating-Incremental" (Alt-Inc) algorithm, which alternates between optimizing cost function and pseudo labels in [334]<br>• Self-supervised contrastive learning along with Transformer encoder-decoder in [335]<br>• Multi-Net [336] |
| 7 | Banking and Finance | Credit Risk / Underwriting Process and Assessment | • Adapting entities extraction (credit risk attributes) from financial documents FIN [238] from the English dataset CoNLL2003 [231].<br>• Adapting models for geography and cultural features to understand the likelihood of default. | • NER DA in [238] |
| 8 | Banking and Finance | Predicting Financial Outcomes | • Sedinkina, Breitkopf, and Schütze [337] mentioned that the use of DA to create a sentiment dictionary outperformed existing methods related to financial outcomes- excess return & volatility. | • Pseudo Labeling: Labels from the source domains are used to train new models from scratch |
| 9 | Banking and Finance | Fraud | • Lebichot *et al.* [338] adapted for credit card fraud behavior (source domain: e-commerce and target domain: face-to-face), which differs across payment systems, countries, and population segments. | • 5 techniques are used<br>1. Baseline-only target data used<br>2. Naïve – using source domain classifier on the target domain<br>3. Feature representation (Imputation) based on [18] |

**TABLE 22.** *(Continued.)* **Industrial applications: DA use-cases across industries.**

| S. No. | Industry | Area / Sub-Area | DA Use-case | DA Technique / Method |
|---|---|---|---|---|
| | | | | 4. Adding additional features of the source to target based on [339]<br>5. Adversarial [49]) |
| 10 | Civil Engineering / Construction | Structural Health Monitoring | • Gardner *et al.* [340] used DA in structural health monitoring (SHM) to understand the structural health of buildings and wind turbines. It is arduous to get labeled data in case of failure as it is very little. Data from a handful of failure occurrences is used in one setting and adapted to a different setting. | • 3 techniques are used<br>1. Transfer Component Analysis (TCA) [15]<br>2. Joint Domain Adaption (JDA) [41]<br>3. Adaptation Regularization based Transfer Learning (ARTL) [341] |
| 11 | Education | Automating Admin Tasks | • Admin tasks like grading, Question, and Answers, differ a lot in hard sciences and soft subjects. A model for one subject can be used for another. | • Ganin and Lempitsky [49] or similar |
| 12 | Healthcare | AI Imaging | • Kouw and Loog [342] suggest that DA can reduce a considerable variation between data sets of CT, MRI, or PET scanners centers where the shift is due to vendor, calibration, mechanical configuration, or acquisition protocol.<br>• CIDA [158] understood the implications of age and other demographic shifts and helped adapt to these shifts in a sleep study. CIDA can predict the stage of sleep in different demographic shifts (age, gender, etc.)<br>• Ren *et al.* [343] mentioned that it is important to evaluate digitized histopathology specimens as they increase the reliability of cancer diagnoses and help understand underlying mechanisms of disease onset. However, it isn't easy to analyze those samples due to the difference in tools and techniques used to take those specimens. DA played a vital role in reducing this domain shift in samples.<br>• Tang *et al.* [344] used Active and Transfer Learning (Active Learning on the target dataset and DA on the source dataset) | • Adversarial discriminative method [345] mentioned in [342]<br>• CIDA [158]<br>• Siamese network architecture with adversarial generative models is very similar to CoGAN [194] in [343]<br>• Used Instance Selection by removing samples from a source domain, used Naïve Bayes classifier and KL Divergence in [344] |
| 13 | Industrial Manufacturing | Digital Twins | • The digital twin is a representation (visual) of a real-world asset. AI simulations on digital twins can be adapted using domain adaptation for real-world assets. Liu, Mauricio, Qi, Peng, & Gryllias [346] used DANN [49] to predict Remaining Useful Life (RUL). | • Adversarial Model - DANN [49] |
| 14 | Industrial Manufacturing | Predictive Maintenance | • Mahyari and Locker [347] used DA to understand the predictive maintenance of robots in an actual setting, the features extracted during the actual setting are different from the trained model. | • Manifold Alignment [348] (Feature Matching or Transformation strategy) |
| 15 | Industrial Manufacturing | Edge Analytics | • Yang *et al.* [349] proposed Mobile DA framework based on the teacher (on a server) – student (on edge device) network and cross-domain distillation. Student network in the new domain (edge) amends feature learning to be domain invariant. | • Novel architecture called Mobile DA framework in [349] |
| 16 | Life Sciences | Drug response | • Mourragui *et al.* [350] used the DA approach to transfer drug response predictors from pre-clinical models (source domain) to the human setting (target domain). | • A novel strategy of Patient Response Estimation Corrected by Interpolation of Subspace Embeddings (PRECISE) in [350] |
| 17 | Remote Sensing | Satellite Imagery | • Deng *et al.* [351] used DA in remote sensing imagery semantic segmentation. They assumed that object classes and the structures in the scene are similar in the two sets. | • An adversarial generative model for DA with a combined segmentation loss and adversarial loss in [351] |
| 18 | Security and Safety | Person Identification/ re-identification | • DA has found fair use in face recognition [352] provided an initial view of the complexities - different face poses, different headgears, different emotions, and different sketches. Sohn *et al.* [296] also provided usage for face recognition. This can be used for forensic sciences to match the sketches (one domain) with photographs (another domain) and look for the same label. Further, another possible use case is identifying a person based on speech in different settings (loudspeaker and phone). | • Adversarial discriminative models variant used in [296] |

**TABLE 22.** *(Continued.)* **Industrial applications: DA use-cases across industries.**

| S. No. | Industry | Area / Sub-Area | DA Use-case | DA Technique / Method |
|---|---|---|---|---|
| 19 | Telecommunications | Software Defined Networks (SDN) | • Latah and Toker [353] mention that DA can be relevant in AI models centrally controlling multiple network systems, where adaptation is required amongst different sub-networks as features can be a little different. | • Ensemble-based methods in [353] |
| 20 | Transportation | Navigation | • Yoo *et al.* [354] used DA with adversarial learning to generate images with realistic textures reducing the effort required and dependence on real labeled data. | • The adversarial generative model with style loss and cycle loss is similar to CycleGAN [63] in [354] |
| 21 | Utilities | Consumption Monitoring | • Liu *et al.* [355] used the Non-intrusive load monitoring (NILM) | • Discrepancy based method - Joint Adaptation Network (JAN) [41] |

proliferated usage in NLP and CV is used in DA (DAF [137], Adversarial Memory Network (AMN) (Attention + DANN + SCL MemNet) [55], Federated domain adaptation [160]). Also, the focus is to include two or more DA techniques together. Wilson and Cook [356] mention the combination of the teacher-student network [84] and AutoDIAL [110], AutoDIAL can replace the student network to understand the degree of adaptation. Similarly, GAN, a data augmentation technique, can replace stochastic data augmentation in [84]. This augmentation of multiple techniques or methods can be useful in multimodal DA.

- **Multi-domain support**: To support multiple domains in DA, techniques or methods are required to deal with larger domain shifts and/or are robust. StarGAN [357] looks at multi-domain image-to-image translation and can be used in multi-domain adaptation.
- **Cross-modal application**: DA techniques or methods primarily developed for one modality (say text) can be used in another modality (say an image). It is observed currently that other than adversarial methods, not many methods are used across modalities.
- **Supporting more real-world scenarios**: DA researchers are looking to support more real-world scenarios. These real-world which are inspired by data (unavailability, label-set difference, etc.) and environmental (restricted, sequential, etc.) limitations. The current research endeavor is to support a larger domain shift in DA when applied to real-world applications. WILDS Datasets [358] provide 10 curated real-world dataset benchmarks having a varied range of domain shifts. Further, DA provides the potential for reinforcement learning applications to learn in a simulated environment and then apply the policy learned to the real-world environment. More industrial applications as part of Industry 4.0 can be supported by DA. For example, IoT devices or edge devices are quite varied, and they are installed in varied environments / used by varied users; this variation provides good ground to use DA.
- **Use of more stable training approaches**: Adversarial feature learning-based approaches are still most utilized by researchers, even though the training at times is unstable in practice and requires careful selection and tuning of parameters. However, pseudo-learning-based approaches (including pseudo-learning based self-training) are being adopted by researchers more and more based on their outperformance and training stability. However, one drawback of pseudo-learning-based approaches indeed is noise in pseudo labels, which can lead to under performance. Focus of researchers are now looking to employ only more confident pseudo-predictions for training. Similarly, the use of mean-teacher strategy is on the rise, as the approach utilities additional regularizations or feature matching strategy which improve the performance.
- **Post-DA over pre-DA strategies**: Post-DA techniques are becoming more common to improve ''fallen'' task accuracy. For example, Saunders and Byrne [355] used Elastic Weight Consolidation (EWC) and lattice-rescoring technique to prop-up the ''fallen'' accuracy (due to catastrophic forgetting during DA). However, pre-DA methods are not much found in the literature. Incorporating pre-DA knowledge of domain gaps arising from either data processing (image processing techniques, text extraction techniques) may lead to a performance increase. One possible way to incorporate this would be to use multi-level constraints in adversarial-based approaches. Further research work undertaken in both Pre and Post DA strategies would improve task accuracies.

**TABLE 23. DA can be applied to a host of industrial use-cases.**

| Industry | Area / Sub-Area | Potential Domain Adaptation Use-case |
|---|---|---|
| Agriculture | Crop Yield | Talaviya *et al.* [359] mentioned AI model adaptation to be used across crops and geographies. DA method would depend on the source and target data |
| Agriculture | Pest infestation | Adapting to pest profile and geography can predict pest infestation better [360]. |
| AI / Software Engineering | Machine Learning Operations (MLOps) | Due to domain shift or concept shift, DA can be employed as a re-training strategy. |
| Fast Moving Consumer Goods (FMCG) | Product Diversity/Bias in Systems | Soap in western countries is liquid, while it is solid in non-western countries [293] |
| Healthcare | Electronic Health Record (EHR) / Risk Prediction | Researchers identify novel connections between seemingly unrelated health aspects. EHR analytics is helpful in stratification and risk-scoring. DA can support deviations based on patients' background (Race, conditions, genetic makeup, previous history) |
| Insurance | Risk Profile | DA can be used for models to adapt to data across geography, demographics, etc. |
| Life Sciences | Drug discovery process / Patient Stratification | The probability of trial success can be improved when the patient data is domain adapted for genetic, geographic, and epidemiologic aspects while doing patient stratification and identification. |
| Media and Entertainment | Metadata Tagging | Inspecting frame by frame and adding objects or tags by AI processes can be made more precise using DA for features related to genre, geographical area, etc. |
| Oil & Gas, Mining | Precision to haul | AI models, when used to mine with precision, reducing environment print and reducing the danger to frontline workers, can further be domain adapted to support the same models across different ores, geography, and environment understanding. |
| Retail | AI-powered data creation | Human fatigue is eliminated by creating AI-powered metadata for each product. This can further be boosted by adapting the model from one product category to another. |
| Retail | AI-powered Guided Discovery | Adapting AI-powered guided discovery from one aspect of retail (say clothes) to another (say shoes) |
| Travel and Hospitality | Operations | Airlines use AI for decision-making when an event occurs (flight delay, weather, restrictions) to reduce the impact on passengers and reduce costs. DA can further boost efficiency when models support multiple geographies/airports. |
| Utilities | Investment / Field Force Planning | AI models for demand and supply forecasting used for one utility (say water) can be used for another utility (say electricity) |

- **Removing bias for specific frameworks:** Just like we see a classification task bias for nearly all DA work, there also exists a research bias for specific frameworks. A case in point is object detection DA, where nearly all the DA strategies focus on Faster RCNN. Other frameworks like YOLO, SSD, and DETR must also be evaluated for DA performance.
- **Solving Industrial use-cases:** DA has the potential to solve many AI industrial use-cases, which are not implemented due to economies of scale in implementation for multiple locations, multiple cultures, multiple demographics, etc., large domain gap is understood in high frequency, etc. Table 23 provides a list of industrial use-cases where DA would lead to enormous benefits for the industry if applied.

## IX. CONCLUSION

There is an imperative need for deep networks to adapt to multiple domains to reduce costs, increase application, and be more human-like - the ultimate aim of artificial intelligence. This paper explores the work done in DA in deep neural networks (also known as deep DA) in multiple data domains (computer vision, NLP, multimodal, speech, time-series), reviews different methods and techniques, and mentions emerging datasets related to DA. This paper focuses on applying DA in more practical settings, in various industries, in the wild, and in real-world scenarios where the DA challenges lie. We believe that research undertaken in mentioned future research frontiers would greatly impact DA and AI as a whole.

## ACKNOWLEDGMENT

The authors would like to thank Dr. Prathosh, an Assistant Professor of the Indian Institute of Science, for the insights and perspective provided on domain adaptation in his course "Advanced Topics on Deep Learning."

## REFERENCES

[1] C. Darwin, *On the Origin of Species by Means of Natural Selection, or Preservation of Favoured Races in the Struggle for Life*. London, U.K.: John Murray, 1859.

[2] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*.

[3] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1–9.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 1–15.

[6] S. J. Pan and Q. Yang, "A survey on transfer learning," in *Proc. IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[7] T. Tommasi, "Tutorial on domain adaptation," in *Proc. ECCV*, 2020, pp. 1–10.

[8] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Proc. NeurIPS*, 2006, pp. 1–8.

[9] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, nos. 1–2, pp. 151–175, May 2010.

[10] H. Zhao, R. T. Des Combes, Z. Kun, and G. Geoffrey, "On learning invariant representations for domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7523–7532.
[11] T. Le, K. Nguyen, N. Ho, H. Bui, and D. Phung, "On deep domain adaptation: Some theoretical understandings," 2018, *arXiv:1811.06199*.
[12] G. Csurka, "Domain adaptation for visual applications: A comprehensive survey," 2017, *arXiv:1702.05374*.
[13] S. J. Pan, J. T. Kwok, and Q. Yang, "Transfer learning via dimensionality reduction," in *Proc. 23rd AAAI Conf. Artif. Intell.*, 2008, pp. 677–682.
[14] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Sch?lkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. 49–57, 2006.
[15] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.
[16] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2006, pp. 120–128.
[17] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *J. Mach. Learn. Res.*, vol. 6, no. 11, pp. 185–1817, 2005.
[18] H. Daumé, "Frustratingly easy domain adaptation," 2009, *arXiv:0907.1815*.
[19] R. Gopalan, R. Li, and R. Chellappa, "Unsupervised adaptation across domain shifts by generating intermediate data representations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2288–2302, Nov. 2014.
[20] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2066–2073.
[21] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2960–2967.
[22] M. Long, G. Ding, J. Wang, J. Sun, Y. Guo, and P. S. Yu, "Transfer sparse coding for robust image representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 407–414.
[23] J. D. Jong. (Oct. 2017). *Transfer Learning: Domain Adaptation by Instance-Reweighting*. Accessed: Sep. 18, 2021. [Online]. Available: https://johanndejong.wordpress.com/2017/10/15/transfer-learning-domain-adaptation-by-instance-reweighting/
[24] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer joint matching for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1410–1417.
[25] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proc. 24th Int. Conf. Mach. Learn.*, Jun. 2007, pp. 1855–1862.
[26] S. Ruder, P. Ghaffari, and J. G. Breslin, "Knowledge adaptation: Teaching to adapt," 2017, *arXiv:1702.02052*.
[27] B. Zadrozny, "Learning and evaluating classifiers under sample selection bias," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, p. 114.
[28] M. Sugiyama, S. Nakajima, H. Kashima, P. Buenau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *Proc. 20th Int. Conf. Neural Inf. Process. Syst.*, 2007, pp. 1–8.
[29] L. Duan, D. Xu, and I. W. Tsang, "Learning with augmented features for heterogeneous domain adaptation," in *Proc. 29th Int. Conf. Int. Conf. Mach. Learn.*, 2012, pp. 1–12.
[30] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," in *Proc. CVPR*, Jun. 2011, pp. 1785–1792.
[31] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Proc. Neural Inf. Process. Syst.*, 2016, pp. 1–15.
[32] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory*, Jul. 1998, pp. 92–100.
[33] S. Ruder, "Neural transfer learning for natural language processing," Ph.D. thesis, School Eng. Inform., Nat. Univ. Ireland, Galway, Ireland, 2019.
[34] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proc. Int. Conf. a Learn. (ICML)*, 2011, pp. 1–11.
[35] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, Oct. 2018.
[36] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 443–450.
[37] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. 30th AAAI Conf. Intell.*, 2016, pp. 1–8.
[38] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4893–4902.
[39] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 97–105.
[40] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1–6.
[41] B. B. Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty, "DeepJDOT: Deep joint distribution optimal transport for unsupervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 447–463.
[42] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy, "Joint distribution optimal transportation for domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–12.
[43] F. Zhuang, X. Cheng, P. Luo, S. J. Pan, and Q. He, "Supervised representation learning: Transfer learning with deep autoencoders," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 1–7.
[44] C.-Y. Lee, T. Batra, M. H. Baig, and D. Ulbricht, "Sliced Wasserstein discrepancy for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10285–10295.
[45] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv:1412.3474*.
[46] A. R. Kashyap, D. Hazarika, M.-Y. Kan, and R. Zimmermann, "Domain divergences: A survey and empirical analysis," 2020, *arXiv:2010.12198*.
[47] Y. Pan, T. Yao, Y. Li, Y. Wang, C.-W. Ngo, and T. Mei, "Transferrable prototypical networks for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2239–2247.
[48] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7167–7176.
[49] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.
[50] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "FCNs in the wild: Pixel-level adversarial and constraint-based adaptation," 2016, *arXiv:1612.02649*.
[51] H. Huang, Q. Huang, and P. Krahenbuhl, "Domain transfer through deep activation matching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 590–605.
[52] Y.-B. Kim, K. Stratos, and D. Kim, "Adversarial adaptation of synthetic or stale data," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1297–1307.
[53] B. Xu, M. Mohtarami, and J. Glass, "Adversarial domain adaptation for stance detection," 2019, *arXiv:1902.02401*.
[54] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–6.
[55] Z. Li, Y. Zhang, Y. Wei, Y. Wu, and Q. Yang, "End-to-end adversarial memory network for cross-domain sentiment classification," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2237–2243.
[56] A. Naik and C. Rose, "Towards open domain event trigger identification using adversarial domain adaptation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 1–7.
[57] M. Yasunaga, J. Kasai, and D. Radev, "Robust multilingual part-of-speech tagging via adversarial training," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, pp. 1–15.
[58] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Proc. NeurIPS*, 2018, pp. 1–13.
[59] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*.

[60] Y. Du, Z. Tan, Q. Chen, X. Zhang, Y. Yao, and C. Wang, "Dual adversarial domain adaptation," 2020, *arXiv:2001.00153*.

[61] H. Rangwani, S. K. Aithal, M. Mishra, A. Jain, and R. V. Babu, "A closer look at smoothness in domain adversarial training," in *Proc. 39th Int. Conf. Mach. Learn. (ICML)*, 2022, pp. 1–22.

[62] G. Shi, C. Feng, L. Huang, B. Zhang, H. Ji, L. Liao, and H. Huang, "Genre separation network with adversarial training for cross-genre relation extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1018–1023.

[63] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[64] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 1989–1998.

[65] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 597–613.

[66] Y. Ziser and R. Reichart, "Neural structural correspondence learning for domain adaptation," in *Proc. 21st Conf. Comput. Natural Lang. Learn.*, 2017, pp. 1–11.

[67] Y. Ziser and R. Reichart, "Pivot based language modeling for improved neural domain adaptation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, pp. 1241–1251.

[68] Y. Ziser and R. Reichart, "Task refinement learning for improved accuracy and stability of unsupervised domain adaptation," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 5895–5906.

[69] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1–10.

[70] S. Sun, H. Shi, and Y. Wu, "A survey of multi-source domain adaptation," *Inf. Fusion*, vol. 24, pp. 84–92, Jul. 2015.

[71] S. Zhao, B. Li, C. Reed, P. Xu, and K. Keutzer, "Multi-source domain adaptation in the deep learning era: A systematic survey," 2020, *arXiv:2002.12169*.

[72] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1406–1415.

[73] S. Zhao, B. Li, X. Yue, Y. Gu, P. Xu, R. Hu, H. Chai, and K. Keutzer, "Multi-source domain adaptation for semantic segmentation," 2019, *arXiv:1910.12181*.

[74] P. Russo, T. Tommasi, and B. Caputo, "Towards multi-source adaptive semantic segmentation," in *Proc. Int. Conf. Image Anal. Process.*, 2019, pp. 292–301.

[75] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," 2017, *arXiv:1705.08045*.

[76] Y. Yang and T. M. Hospedales, "A unified perspective on multi-domain and multi-task learning," 2014, *arXiv:1412.7489*.

[77] H. Guo, R. Pasunuru, and M. Bansal, "Multi-source domain adaptation for text classification via DistanceNet-bandits," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 7830–7838.

[78] J. Guo, D. J. Shah, and R. Barzilay, "Multi-source domain adaptation with mixture of experts," 2018, *arXiv:1809.02256*.

[79] J. Zhu, N. Chen, and C. Shen, "A new multiple source domain adaptation fault diagnosis method between different rotating machines," *IEEE Trans. Ind. Informat.*, vol. 17, no. 7, pp. 4788–4797, Jul. 2021.

[80] Y. Xia, C. Shen, D. Wang, Y. Shen, W. Huang, and Z. Zhu, "Moment matching-based intraclass multisource domain adaptation network for bearing fault diagnosis," *Mech. Syst. Signal Process.*, vol. 168, Apr. 2022, Art. no. 108697.

[81] B. Gholami, P. Sahu, O. Rudovic, K. Bousmalis, and V. Pavlovic, "Unsupervised multi-target domain adaptation: An information theoretic approach," *IEEE Trans. Image Process.*, vol. 29, pp. 3993–4002, 2020.

[82] T. Isobe, X. Jia, S. Chen, J. He, Y. Shi, J. Liu, H. Lu, and S. Wang, "Multi-target domain adaptation with collaborative consistency learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8187–8196.

[83] L. T. Nguyen-Meidine, A. Belal, M. Kiran, J. Dolz, L.-A. Blais-Morin, and E. Granger, "Unsupervised multi-target domain adaptation through knowledge distillation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1339–1347.

[84] G. French, M. Mackiewicz, and M. Fisher, "Self-ensembling for visual domain adaptation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–20.

[85] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," 2017, *arXiv:1703.01780*.

[86] S. Desai, B. Sinno, A. Rosenfeld, and J. J. Li, "Adaptive ensembling: Unsupervised domain adaptation for political document analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 1–13.

[87] K. Saito, Y. Ushiku, and T. Harada, "Asymmetric tri-training for unsupervised domain adaptation," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 2988–2997.

[88] Y. Zou, Z. Yu, B. V. K. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 289–305.

[89] F. Qi, X. Yang, and C. Xu, "A unified framework for multimodal domain adaptation," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 429–437.

[90] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," 2016, *arXiv:1606.01847*.

[91] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," 2016, *arXiv:1610.04325*.

[92] H. Ben-younes, R. Cadene, M. Cord, and N. Thome, "MUTAN: Multimodal tucker fusion for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2612–2620.

[93] H. Zhang, S. Qian, Q. Fang, and C. Xu, "Multimodal disentangled domain adaption for social media event rumor detection," *IEEE Trans. Multimedia*, vol. 23, pp. 4441–4454, 2020.

[94] J. Munro and D. Damen, "Multi-modal domain adaptation for fine-grained action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 122–132.

[95] H. Li, Y. Kim, C.-H. Kuo, and S. Narayanan, "Acted vs. improvised: Domain adaptation for elicitation approaches in audio-visual emotion recognition," 2021, *arXiv:2104.01978*.

[96] X. Shu, G.-J. Qi, J. Tang, and J. Wang, "Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 35–44.

[97] Z. Ding, M. Shao, and Y. Fu, "Missing modality transfer learning via latent low-rank constraint," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4322–4334, Nov. 2015.

[98] C. Athanasiadis, E. Hortal, and S. Asteriadis, "Audio–visual domain adaptation using conditional semi-supervised generative adversarial networks," *Neurocomputing*, vol. 397, pp. 331–344, Jul. 2020.

[99] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *J. Mach. Learn. Res.*, vol. 9, no. 3, pp. 1–51, 2008.

[100] S. Seo, S. Na, and J. Kim, "HMTL: Heterogeneous modality transfer learning for audio-visual sentiment analysis," *IEEE Access*, vol. 8, pp. 140426–140437, 2020.

[101] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," in *Proc. ICLR*, 2018, pp. 1–11.

[102] A. Søgaard, S. Ruder, and I. Vulić, "On the limitations of unsupervised bilingual dictionary induction," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1–11.

[103] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2019, *arXiv:1911.02116*.

[104] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1060–1069.

[105] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5907–5915.

[106] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[107] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[108] W.-G. Chang, T. You, S. Seo, S. Kwak, and B. Han, "Domain-specific batch normalization for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7354–7362.
[109] Y. Li, N. Wang, J. Shi, X. Hou, and J. Liu, "Adaptive batch normalization for practical domain adaptation," *Pattern Recognit.*, vol. 80, pp. 109–117, Aug. 2018.
[110] F. M. Carlucci, L. Porzi, B. Caputo, E. Ricci, and S. R. Bulo, "AutoDIAL: Automatic domain alignment layers," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5067–5075.
[111] S. Roy, A. Siarohin, E. Sangineto, S. R. Bulo, N. Sebe, and E. Ricci, "Unsupervised domain adaptation using feature-whitening and consensus loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9471–9480.
[112] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2229–2238.
[113] J. Xu, L. Xiao, and A. M. López, "Self-supervised domain adaptation for computer vision tasks," *IEEE Access*, vol. 7, pp. 156694–156706, 2019.
[114] D. Kim, K. Saito, T.-H. Oh, B. A. Plummer, S. Sclaroff, and K. Saenko, "Cross-domain self-supervised learning for domain adaptation with few source labels," 2020, *arXiv:2003.08264*.
[115] A. Ramponi and B. Plank, "Neural unsupervised domain adaptation in NLP—A survey," 2020, *arXiv:2006.00632*.
[116] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[117] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 1–6.
[118] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Sep. 2019.
[119] X. Han and J. Eisenstein, "Unsupervised domain adaptation of contextualized embeddings for sequence labeling," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 1–12.
[120] A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov, "XNLI: Evaluating cross-lingual sentence representations," 2018, *arXiv:1809.05053*.
[121] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1–12.
[122] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing LSTM language models," 2017, *arXiv:1708.02182*.
[123] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1–12.
[124] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 1–19.
[125] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[126] J. Phang, T. Févry, and S. R. Bowman, "Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks," 2018, *arXiv:1811.01088*.
[127] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, *arXiv:1802.05365*.
[128] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. (2018). *Improving Language Understanding by Generative Pre-Training*. Accessed: Sep. 18, 2021. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
[129] Y. Zhou and S. Goldman, "Democratic co-learning," in *Proc. 16th IEEE Int. Conf. Tools Artif. Intell.*, 2004, pp. 594–602.
[130] S. Ruder and B. Plank, "Strong baselines for neural semi-supervised learning under domain shift," in *Proc. ACL*, 2018, pp. 1–11.
[131] P. Gimeno, D. Ribas, A. Ortega, A. Miguel, and E. Lleida, "Unsupervised adaptation of deep speech activity detection models to unseen domains," *Appl. Sci.*, vol. 12, no. 4, p. 1832, Feb. 2022.
[132] A. P. Prathosh and A. G. Ramakrishnan, "Unsupervised domain adaptation schemes for building ASR in low-resource languages," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2021, pp. 342–349.
[133] J.-H. Park, M. Oh, and H.-M. Park, "Unsupervised speech domain adaptation based on disentangled representation learning for robust speech recognition," 2019, *arXiv:1904.06086*.
[134] T. Asami, R. Masumura, Y. Yamaguchi, H. Masataki, and Y. Aono, "Domain adaptation of DNN acoustic models using knowledge distillation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 5185–5189.
[135] Z. Meng, J. Li, Y. Gaur, and Y. Gong, "Domain adaptation via teacher-student learning for end-to-end speech recognition," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2019, pp. 268–275.
[136] X.-L. Zhang, "Unsupervised domain adaptation for deep neural network based voice activity detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 6864–6868.
[137] X. Jin, Y. Park, D. C. Maddix, H. Wang, and Y. Wang, "Domain adaptation for time series forecasting via attention sharing," 2021, *arXiv:2102.06828*.
[138] Y. Shi, X. Ying, and J. Yang, "Deep unsupervised domain adaptation with time series sensor data: A survey," *Sensors*, vol. 22, no. 15, p. 5507, Jul. 2022.
[139] Q. Wang, C. Taal, and O. Fink, "Integrating expert knowledge with domain adaptation for unsupervised fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.
[140] A. R. Sanabria, F. Zambonelli, S. Dobson, and J. Ye, "ContrasGAN: Unsupervised domain adaptation in human activity recognition via adversarial and contrastive learning," *Pervas. Mobile Comput.*, vol. 78, Dec. 2021, Art. no. 101477.
[141] B. Yang, Q. Li, L. Chen, C. Shen, and S. Natarajan, "Bearing fault diagnosis based on multilayer domain adaptation," *Shock Vib.*, vol. 2020, pp. 1–11, Sep. 2020.
[142] S. Motiian, Q. Jones, S. M. Iranmanesh, and G. Doretto, "Few-shot adversarial domain adaptation," 2017, *arXiv:1711.02536*.
[143] A. Zhao, M. Ding, Z. Lu, T. Xiang, Y. Niu, J. Guan, and J.-R. Wen, "Domain-adaptive few-shot learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1390–1399.
[144] X. Yue, Z. Zheng, S. Zhang, Y. Gao, T. Darrell, K. Keutzer, and A. S. Vincentelli, "Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13834–13844.
[145] K. C. Peng, Z. Wu, and J. Ernst, "Zero-shot deep domain adaptation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 764–781.
[146] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Unsupervised domain adaptation for zero-shot learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2452–2460.
[147] K. Saito, S. Yamamoto, Y. Ushiku, and T. Harada, "Open set domain adaptation by backpropagation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 153–168.
[148] H. Liu, Z. Cao, M. Long, J. Wang, and Q. Yang, "Separate to adapt: Open set domain adaptation via progressive separation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2927–2936.
[149] X. Peng, B. Usman, K. Saito, N. Kaushik, J. Hoffman, and K. Saenko, "Syn2Real: A new benchmark for synthetic-to-real visual domain adaptation," 2018, *arXiv:1806.09755*.
[150] Y. Pan, T. Yao, Y. Li, C.-W. Ngo, and T. Mei, "Exploring category-agnostic clusters for open-set domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13867–13875.
[151] Z. Cao, L. Ma, M. Long, and J. Wang, "Partial adversarial domain adaptation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 135–150.
[152] Z. Cao, K. You, M. Long, J. Wang, and Q. Yang, "Learning to transfer examples for partial domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2985–2994.

[153] K. You, M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Universal domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2720–2729.

[154] J. N. Kundu, N. Venkat, M. V. Rahul, and R. V. Babu, "Universal source-free domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4544–4553.

[155] M. Mancini, H. Karaoguz, E. Ricci, P. Jensfelt, and B. Caputo, "Kitting in the wild through online domain adaptation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1103–1109.

[156] M. Mancini, S. R. Bulo, B. Caputo, and E. Ricci, "AdaGraph: Unifying predictive and continuous domain adaptation through graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6568–6577.

[157] A. Bobu, E. Tzeng, J. Hoffman, and T. Darrell, "Adapting to continuously shifting domains," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, 2018, pp. 1–4.

[158] H. Wang, H. He, and D. Katabi, "Continuously indexed domain adaptation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1–14.

[159] Z. Liu, Z. Miao, X. Pan, X. Zhan, D. Lin, S. X. Yu, and B. Gong, "Open compound domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1–10.

[160] X. Peng, Z. Huang, Y. Zhu, and K. Saenko, "Federated adversarial domain adaptation," 2019, *arXiv:1911.02054*.

[161] S. Bucci, A. D'Innocente, Y. Liao, F. M. Carlucci, B. Caputo, and T. Tommasi, "Self-supervised learning across domains," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5516–5528, Sep. 2022.

[162] S. Bucci, A. D'Innocente, and T. Tommasi, "Tackling partial domain adaptation with self-supervision," in *Proc. Int. Conf. Image Anal. Process.*, 2019, pp. 70–81.

[163] S. Bucci, M. R. Loghmani, and T. Tommasi, "On the effectiveness of image rotation for open set domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 422–438.

[164] D. Li and T. Hospedales, "Online meta-learning for multi-source and semi-supervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 382–403.

[165] R. Ribeiro, A. Abad, and J. Lopes, "Domain adaptation in dialogue systems using transfer and meta-learning," 2021, *arXiv:2102.11146*.

[166] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[167] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "MetaReg: Towards domain generalization using meta-regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1–8.

[168] J.-C. Su, Y.-H. Tsai, K. Sohn, B. Liu, S. Maji, and M. Chandraker, "Active adversarial domain adaptation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 739–748.

[169] V. Prabhu, A. Chandrasekaran, K. Saenko, and J. Hoffman, "Active domain adaptation via clustering uncertainty-weighted embeddings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8505–8514.

[170] Y. Kim and C. Kim, "Semi-supervised domain adaptation via selective pseudo labeling and progressive self-training," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 1059–1066.

[171] K. Wang, N. Thakur, N. Reimers, and I. Gurevych, "GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2022, pp. 1–16.

[172] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[173] J. Wang, J. Chen, J. Lin, L. Sigal, and C. W. de Silva, "Discriminative feature alignment: Improving transferability of unsupervised domain adaptation by Gaussian-guided latent alignment," *Pattern Recognit.*, vol. 116, Aug. 2021, Art. no. 107943.

[174] S. Xie, Z. Zheng, L. Chen, and C. Chen, "Learning semantic representations for unsupervised domain adaptation," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 5423–5432.

[175] C. Chen, W. Xie, W. Huang, Y. Rong, X. Ding, Y. Huang, T. Xu, and J. Huang, "Progressive feature alignment for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 627–636.

[176] S. Lee, S. Cho, and S. Im, "DRANet: Disentangling representation and adaptation networks for unsupervised cross-domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15252–15261.

[177] J. J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 5, pp. 550–554, May 1994.

[178] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3723–3732.

[179] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *Proc. NIPS Workshop Deep Learn. Unsupervised*, 2011, pp. 1–9.

[180] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 213–226.

[181] J. Na, H. Jung, H. J. Chang, and W. Hwang, "FixBi: Bridging domain spaces for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1094–1103.

[182] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, "Generate to adapt: Aligning domains using generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8503–8512.

[183] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5018–5027.

[184] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation," 2020, *arXiv:2002.08546*.

[185] Q. Wang and T. P. Breckon, "Unsupervised domain adaptation via structured prediction based selective pseudo-labeling," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 6243–6250.

[186] B. Gong, K. Grauman, and F. Sha, "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 222–230.

[187] H.-Z. Feng, Z. You, M. Chen, T. Zhang, M. Zhu, F. Wu, C. Wu, and W. Chen, "KD3A: Unsupervised multi-source decentralized domain adaptation via knowledge distillation," 2020, *arXiv:2011.09757*.

[188] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[189] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke, "YouTube-BoundingBoxes: A large high-precision human-annotated data set for object detection in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5296–5305.

[190] H. Tang and K. Jia, "Discriminative adversarial domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 5940–5947.

[191] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, and T. Yao, "Exploring object relation in mean teacher for cross-domain detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11457–11466.

[192] X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang, and K. Saenko, "VisDA: The visual domain adaptation challenge," 2017, *arXiv:1710.06924*.

[193] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[194] L. Ming-Yu and T. Oncel, "Coupled generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–13.

[195] Q. Zhou, Q. Gu, J. Pang, X. Lu, and L. Ma, "Self-adversarial disentangling for specific domain adaptation," 2021, *arXiv:2108.03553*.

[196] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 973–992, 2018.

[197] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The German traffic sign recognition benchmark: A multi-class classification competition," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2011, pp. 1453–1460.

[198] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 102–118.
[199] P. Zhang, B. Zhang, T. Zhang, D. Chen, Y. Wang, and F. Wen, "Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12414–12424.
[200] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3234–3243.
[201] N. Araslanov and S. Roth, "Self-supervised augmentation consistency for adapting semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15384–15394.
[202] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C.-F. Wang, and M. Sun, "No more discrimination: Cross city adaptation of road scene segmenters," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1992–2001.
[203] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.
[204] M.-H. Chen, Z. Kira, G. Alregib, J. Yoo, R. Chen, and J. Zheng, "Temporal attentive alignment for large-scale video domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6321–6330.
[205] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.
[206] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 548–562.
[207] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4511–4520.
[208] Z. Guo, Z. Yuan, C. Zhang, W. Chi, Y. Ling, and S. Zhang, "Domain adaptation gaze estimation by embedding with prediction consistency," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 1–16.
[209] K. A. F. Mora, F. Monay, and J.-M. Odobez, "EYEDIAP: A database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras," in *Proc. Symp. Eye Tracking Res. Appl.*, Mar. 2014, pp. 255–258.
[210] P. Jiang and S. Saripalli, "LiDARNet: A boundary-aware domain adaptation model for point cloud semantic segmentation," 2020, *arXiv:2003.01174*.
[211] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9297–9307.
[212] Y. Pan, B. Gao, J. Mei, S. Geng, C. Li, and H. Zhao, "SemanticPOSS: A point cloud dataset with large quantity of dynamic instances," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 687–693.
[213] A. Carballo, J. Lambert, A. Monrroy, D. Wong, P. Narksri, Y. Kitsukawa, E. Takeuchi, S. Kato, and K. Takeda, "LIBRE: The multiple 3D LiDAR dataset," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 1094–1101.
[214] L. T. Triess, M. Dreissig, C. B. Rist, and J. M. Zollner, "A survey on deep domain adaptation for LiDAR perception," in *Proc. IEEE Intell. Vehicles Symp. Workshops (IV Workshops)*, Jul. 2021, pp. 350–357.
[215] *Waymo Open Dataset*, Waymo, Mountain View, CA, USA, Jun. 2022.
[216] B. Caine, R. Roelofs, V. Vasudevan, J. Ngiam, Y. Chai, Z. Chen, and J. Shlens, "Pseudo-labeling for scalable 3D object detection," 2021, *arXiv:2103.02093*.
[217] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *Proc. 25th Int. Conf. World Wide Web*, Apr. 2016, pp. 1–11.
[218] Q. Nguyen. (2015). *Skytrax Reviews*. Accessed: Sep. 18, 2021. [Online]. Available: https://github.com/quankiquanki/skytrax-reviews-dataset
[219] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending against neural fake news," 2019, *arXiv:1905.12616*.
[220] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2011, pp. 142–150.
[221] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A large-scale dataset for fact extraction and verification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, pp. 1–20.
[222] M. Davies, "The corpus of contemporary American English: 450 million words, 1990-present," Brigham Young Univ., Provo, UT, USA, Tech. Rep., 2008.
[223] E. Sandhu. (2008). *The New York Times Annotated Corpus LDC2008T19 Linguistic Data Consortium*. Accessed: Sep. 18, 2021. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2008T19
[224] Y. Wu, D. Bamman, and S. Russell, "Adversarial training for relation extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, 2017, pp. 1778–1783.
[225] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of English: The Penn Treebank," *Comput. Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.
[226] J. Nivre, Ž. Agić, M. J. Aranzabe, M. Asahara, A. Atutxa, M. Ballesteros, J. Bauer, K. Bengoetxea, R. A. Bhat, C. Bosco, S. Bowman, G. G. A. Celano and M. Connor. (Nov. 15, 2015). *Universal Dependencies*. Accessed: Sep. 18, 2021. [Online]. Available: http://hdl.handle.net/11234/1-1548
[227] S. Petrov and R. McDonald, "Overview of the 2012 shared task on parsing the web," Google, Tech. Rep., 2012.
[228] C. Galves, "The Tycho Brahe corpus of historical Portuguese: Methodology and results," *Linguistic Variation*, vol. 18, no. 1, pp. 49–73, 2018.
[229] Y. Yang and J. Eisenstein, "Fast easy unsupervised domain adaptation with marginalized structured dropout," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 538–544.
[230] A. Kroch, B. Santorini, and L. Delfs. (2004). *The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)*. Accessed: Sep. 18, 2021. [Online]. Available: http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCEME-RELEASE-3
[231] E. F. T. K. Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," 2003, *arXiv:cs/0306050*.
[232] C. Jia, X. Liang, and Y. Zhang, "Cross-domain NER using Cross-domain language modeling," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2464–2474.
[233] T.-T. Vu, D. Phung, and G. Haffari, "Effective unsupervised domain adaptation with adversarially trained language models," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 1–11.
[234] B. Strauss, B. Toma, A. Ritter, M. D. Marneffe, and W. Xu, "Results of the WNUT16 named entity recognition shared task," in *Proc. 2nd Workshop Noisy User-Generated Text (WNUT)*, 2016, pp. 1–7.
[235] D. Bamman, S. Popat, and S. Shen, "An annotated dataset of literary entities," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Minneapolis, MI, USA, 2019, pp. 2138–2144.
[236] J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo, "The TIMEBANK corpus," in *Proc. Corpus Linguistics Conf.*, 2003, p. 40.
[237] G. Crichton, S. Pyysalo, B. Chiu, and A. Korhonen, "A neural network multi-task learning approach to biomedical named entity recognition," *BMC Bioinf.*, vol. 18, no. 1, pp. 1–14, Dec. 2017.
[238] J. C. S. Alvarado, K. Verspoor, and T. Baldwin, "Domain adaption of named entity recognition to support credit risk assessment," in *Proc. Australas. Lang. Technol. Assoc. Workshop*, 2015, pp. 84–90.
[239] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier, "Introduction to the bio-entity recognition task at JNLPBA," in *Proc. Int. Joint Workshop Natural Lang. Process. Biomed. Appl.*, 2004, pp. 70–75.
[240] N. Poerner, U. Waltinger, and H. Schütze, "Inexpensive domain adaptation of pretrained language models: Case studies on biomedical NER and COVID-19 QA," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2020, pp. 1–8.
[241] L. Smith et al., "Overview of BioCreative II gene mention recognition," *Genome Biol.*, vol. 9, no. 2, pp. 1–19, 2008.

[242] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu, "BioCreative V CDR task corpus: A resource for chemical disease relation extraction," *Database*, vol. 2016, Jan. 2016, Art. no. baw068.

[243] R. I. Doğan, R. Leaman, and Z. Lu, "NCBI disease corpus: A resource for disease name recognition and concept normalization," *J. Biomed. Informat.*, vol. 47, pp. 1–10, Feb. 2014.

[244] M. Krallinger, O. Rabal, F. Leitner, M. Vazquez, D. Salgado, Z. Lu, R. Leaman, Y. Lu, D. Ji, and D. M. Lowe, "The CHEMDNER corpus of chemicals and drugs and its annotation principles," *J. Cheminform.*, vol. 7, no. 1, pp. 1–17, 2015.

[245] M. Gerner, G. Nenadic, and C. M. Bergman, "LINNAEUS: A species name identification system for biomedical literature," *BMC Bioinf.*, vol. 11, no. 1, pp. 1–17, Dec. 2010.

[246] C. Walker, S. Strassel, J. Medero, and K. Maeda. (Feb. 15, 2006). *ACE 2005 Multilingual Training Corpus*. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2006T06

[247] L. Fu, T. H. Nguyen, B. Min, and R. Grishman, "Domain adaptation for relation extraction with domain adversarial neural network," in *Proc. 8th Int. Joint Conf. Natural Lang. Process.*, 2017, pp. 425–429.

[248] S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen, and T. Salakoski, "BioInfer: A corpus for information extraction in the biomedical domain," *BMC Bioinf.*, vol. 8, no. 1, pp. 1–24, Dec. 2007.

[249] A. Rios, R. Kavuluru, and Z. Lu, "Generalizing biomedical relation classification with neural adversarial domain adaptation," *Bioinformatics*, vol. 34, no. 17, pp. 2973–2981, Sep. 2018.

[250] R. Bunescu, R. Ge, R. J. Kate, E. M. Marcotte, R. J. Mooney, A. K. Ramani, and Y. W. Wong, "Comparative experiments on learning information extractors for proteins and their interactions," *Artif. Intell. Med.*, vol. 33, no. 2, pp. 139–155, Feb. 2005.

[251] I. Segura-Bedmar, P. Martínez, and M. Herrero-Zazo, "Lessons learnt from the DDIExtraction-2013 shared task," *J. Biomed. Informat.*, vol. 51, pp. 152–164, Oct. 2014.

[252] A. Liu, S. Soderland, J. Bragg, C. H. Lin, X. Ling, and D. S. Weld, "Effective crowd annotation for relation extraction," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 897–906.

[253] Microsoft Corporation. (Jun. 2019). *Microsoft Research Open Data (Cortana Dataset)*. Accessed: Sep. 18, 2021. [Online]. Available: https://msropendata.com/datasets/1cc496ec-aaff-4576-b4bc-4a65798fa907

[254] D. Pearce and H.-G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. 6th Int. Conf. Spoken Lang. Process.*, Oct. 2000, pp. 1–8.

[255] S. Sun, B. Zhang, L. Xie, and Y. Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, vol. 257, pp. 79–87, Sep. 2017.

[256] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. 2nd Int. Conf. Spoken Lang. Process. (ICSLP)*, Oct. 1992, pp. 1–6.

[257] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.

[258] D. Iskra, B. Grosskopf, K. Marasek, H. V. D. Heuvel, F. Diehl, and A. Kiessling, "SPEECON—Speech databases for consumer devices: Database specification and validation," in *Proc. LREC*, 2002, pp. 1–6.

[259] P. Denisov, N. T. Vu, and M. F. Font, "Unsupervised domain adaptation by adversarial learning for robust speech recognition," in *Proc. Speech Commun.; 13th ITG-Symp.*, 2018, pp. 1–15.

[260] P. Price et al., "Resource management RM1 2.0 LDC93S3B, Web download," Linguistic Data Consortium, Univ. Pennsylvania, Philadelphia, PA, USA, 1993. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S3B

[261] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, Apr. 1995.

[262] J. H. L. Hansen, A. Sangwan, A. Joglekar, A. E. Bulut, L. Kaushik, and C. Yu, "Fearless steps: Apollo-11 corpus advancements for speech technologies from earth to the moon," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 1–5.

[263] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W.-H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, pp. 1–9, May 2016.

[264] S. Purushotham, W. Carvalho, T. Nilanon, and Y. Liu, "Variational recurrent adversarial deep domain adaptation," in *Proc. ICLR*, 2017, pp. 1–15.

[265] R. G. Khemani, D. Conti, T. A. Alonzo, R. D. Bart, and C. J. L. Newth, "Effect of tidal volume in children with acute hypoxemic respiratory failure," *Intensive Care Med.*, vol. 35, no. 8, pp. 1428–1437, Aug. 2009.

[266] A. Jain, H. S. Koppula, S. Soh, B. Raghavan, A. Singh, and A. Saxena, "Brain4Cars: Car that knows before you do via sensory-fusion deep learning architecture," 2016, *arXiv:1601.00740*.

[267] M. Tonutti, E. Ruffaldi, A. Cattaneo, and C. A. Avizzano, "Robust and subject-independent driving manoeuvre anticipation through domain-adversarial recurrent neural networks," *Robot. Auto. Syst.*, vol. 115, pp. 162–173, May 2019.

[268] R. Cai, J. Chen, Z. Li, W. Chen, K. Zhang, J. Ye, Z. Li, X. Yang, and Z. Zhang, "Time series domain adaptation via sparse associative structure alignment," in *Proc. 35th AAAI Conf. Artif. Intell. (AAAI)*, 2021, pp. 6859–6867.

[269] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, "Forecasting fine-grained air quality based on big data," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2015, pp. 2267–2276.

[270] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[271] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia-Mag.*, vol. 19, no. 3, pp. 34–41, Jul. 2012.

[272] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, Jan. 2017.

[273] D. Huang, J. Sun, and Y. Wang, "The BUAA-VisNir face database instructions," Beihang Univ., Beijing, China, Tech. Rep., IRIP-TR-12-FR-001, 2012.

[274] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, Aug. 2011.

[275] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003.

[276] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[277] J.-M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders, "The Amsterdam library of object images," *Int. J. Comput. Vis.*, vol. 61, no. 1, pp. 103–112, Jan. 2005.

[278] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (COIL-100)," Dept. Comput. Sci., Columbia Univ., New York, NY, USA, Tech. Rep. CUCS-006-96, 1996.

[279] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Oct. 2014.

[280] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391.

[281] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016, *arXiv:1606.06259*.

[282] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from national university of Singapore," in *Proc. ACM Int. Conf. Image Video Retr.*, Jul. 2009, pp. 1–9.

[283] A. Zubiaga, M. Liakata, and R. Procter, "Exploiting context for rumour detection in social media," in *Social Informatics*. Cham, Switzerland: Springer, 2017, pp. 109–123.

[284] E. Kochkina, M. Liakata, and A. Zubiaga, "All-in-one: Multi-task learning for rumour verification," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 1–12.

[285] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, and W. Price, "Scaling egocentric vision: The EPIC-kitchens dataset," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 720–736.

[286] J. Liang, Y. Wang, D. Hu, R. He, and J. Feng, "A balanced and uncertainty-aware approach for partial domain adaptation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 123–140.

[287] F. Qiao, L. Zhao, and X. Peng, "Learning to learn single domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12556–12565.

[288] T. M. H. Hsu, W. Y. Chen, C.-A. Hou, Y.-H.-H. Tsai, Y.-R. Yeh, and Y.-C.-F. Wang, "Unsupervised domain adaptation with imbalanced cross-domain data," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4121–4129.

[289] T. Gebru, J. Hoffman, and L. Fei-Fei, "Fine-grained recognition in the wild: A multi-task domain adaptation approach," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1349–1358.

[290] S. Reddy, D. Chen, and C. D. Manning, "CoQA: A conversational question answering challenge," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 249–266, Nov. 2019.

[291] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2827–2836.

[292] D. Saunders and B. Byrne, "Reducing gender bias in neural machine translation as a domain adaptation problem," 2020, *arXiv:2004.04498*.

[293] T. DeVries, I. Misra, C. Wang, and L. V. D. Maaten, "Does object recognition work for everyone?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 52–59.

[294] G. Csurka, R. Volpi, and B. Chidlovskii, "Unsupervised domain adaptation for semantic image segmentation: A comprehensive survey," 2021, *arXiv:2112.03241*.

[295] P. Oza, V. A. Sindagi, and V. M. Patel, "Unsupervised domain adaptation of object detectors: A survey," 2021, *arXiv:2105.13502*.

[296] K. Sohn, S. Liu, G. Zhong, X. Yu, M.-H. Yang, and M. Chandraker, "Unsupervised domain adaptation for face recognition in unlabeled videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3210–3218.

[297] H.-K. Hsu, C.-H. Yao, Y.-H. Tsai, W.-C. Hung, H.-Y. Tseng, M. Singh, and M.-H. Yang, "Progressive domain adaptation for object detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 749–757.

[298] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3339–3348.

[299] W. Xu, J. He, H. L. Zhang, B. Mao, and J. Cao, "Real-time target detection and recognition with deep convolutional networks for intelligent visual surveillance," in *Proc. 9th Int. Conf. Utility Cloud Comput.*, Dec. 2016, pp. 321–326.

[300] J. Zhang, J. Huang, Z. Luo, G. Zhang, and S. Lu, "DA-DETR: Domain adaptive detection transformer by hybrid attention," 2021, *arXiv:2103.17084*.

[301] S. Sankaranarayanan, Y. Balaji, A. Jain, S. N. Lim, and R. Chellappa, "Learning from synthetic data: Addressing domain shift for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3752–3761.

[302] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1–7.

[303] J. N. Kundu, P. K. Uppala, A. Pahuja, and R. V. Babu, "AdaDepth: Unsupervised content congruent adaptation for depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2656–2665.

[304] T.-H. Chen, Y.-H. Liao, C.-Y. Chuang, W.-T. Hsu, J. Fu, and M. Sun, "Show, adapt and tell: Adversarial training of cross-domain image captioner," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 521–530.

[305] W. Zhao, W. Xu, M. Yang, J. Ye, Z. Zhao, Y. Feng, and Y. Qiao, "Dual learning for cross-domain image captioning," in *Proc. ACM Conf. Inf. Knowl. Manage.*, Nov. 2017, pp. 29–38.

[306] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully convolutional localization networks for dense captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4565–4574.

[307] Z. Yang, J. Hu, R. Salakhutdinov, and W. W. Cohen, "Semi-supervised QA with generative domain-adaptive nets," 2017, *arXiv:1702.02206*.

[308] D. Britz, Q. Le, and R. Pryzant, "Effective domain mixing for neural machine translation," in *Proc. 2nd Conf. Mach. Transl.*, 2017, pp. 118–126.

[309] B. Chen, C. Cherry, G. Foster, and S. Larkin, "Cost weighting for neural machine translation domain adaptation," in *Proc. 1st Workshop Neural Mach. Transl.*, 2017, pp. 40–46.

[310] C. Chu and R. Wang, "A survey of domain adaptation for neural machine translation," in *Proc. 27th Int. Conf. Comput. Linguistics*, Sante Fe, NM, USA, 2018, pp. 1–16.

[311] W.-J. Ko, G. Durrett, and J. J. Li, "Domain agnostic real-valued specificity prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6610–6617.

[312] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4889–4893.

[313] S. Khurana, N. Moritz, T. Hori, and J. L. Roux, "Unsupervised domain adaptation for speech recognition via uncertainty driven self-training," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6553–6557.

[314] E. Hosseini-Asl, Y. Zhou, C. Xiong, and R. Socher, "Augmented cyclic adversarial learning for low resource domain adaptation," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–15.

[315] L. Samarakoon, B. Mak, and A. Y. S. Lam, "Domain adaptation of end-to-end speech recognition in low-resource settings," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 382–388.

[316] H. Zhao, Z. Zhu, J. Hu, A. Coates, and G. Gordon, "Principled hybrids of generative and discriminative domain adaptation," 2017, *arXiv:1705.09011*.

[317] X. Zhang, J. Wang, N. Cheng, and J. Xiao, "TDASS: Target domain adaptation speech synthesis framework for multi-speaker low-resource TTS," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–7.

[318] S. Mavaddaty, S. M. Ahadi, and S. Seyedin, "A novel speech enhancement method by learnable sparse and low-rank decomposition and domain adaptation," *Speech Commun.*, vol. 76, pp. 42–60, Feb. 2016.

[319] C. Chen, Y. Miao, C. X. Lu, L. Xie, P. Blunsom, A. Markham, and N. Trigoni, "MotionTransformer: Transferring neural inertial tracking between domains," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8009–8016.

[320] A. Hussein and H. Hajj, "Domain adaptation with representation learning and nonlinear relation for time series," *ACM Trans. Internet Things*, vol. 3, no. 2, pp. 1–26, May 2022.

[321] F. Ott, D. Rügamer, L. Heublein, B. Bischl, and C. Mutschler, "Domain adaptation for time-series classification to mitigate covariate shift," 2022, *arXiv:2204.03342*.

[322] G. Wilson, J. R. Doppa, and D. J. Cook, "Multi-source deep domain adaptation with weak supervision for time-series sensor data," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 1768–1778.

[323] B. Lucas, C. Pelletier, D. Schmidt, G. I. Webb, and F. Petitjean, "Unsupervised domain adaptation techniques for classification of satellite image time series," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep. 2020, pp. 1074–1077.

[324] X. Jin, Y. Park, D. C. Maddix, H. Wang, and Y. Wang, "Domain adaptation for time series forecasting via attention sharing," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 10280–10297.

[325] T. Li, X. Chen, S. Zhang, Z. Dong, and K. Keutzer, "Cross-domain sentiment classification with contrastive learning and mutual information maximization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 8203–8207.

[326] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić, "MultiWOZ—A large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling," 2018, *arXiv:1810.00278*.

[327] W.-N. Zhang, Q. Zhu, Y. Wang, Y. Zhao, and T. Liu, "Neural personalized response generation as domain adaptation," *World Wide Web*, vol. 22, pp. 1427–1446, Jul. 2019.

[328] M. Yang, W. Tu, Q. Qu, Z. Zhao, X. Chen, and J. Zhu, "Personalized response generation by dual-learning based domain adaptation," *Neural Netw.*, vol. 103, pp. 72–82, Jul. 2018.

[329] M. Azamfar, X. Li, and J. Lee, "Deep learning-based domain adaptation method for fault diagnosis in semiconductor manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 33, no. 3, pp. 445–453, Aug. 2020.
[330] M. Thota, S. Kollias, M. Swainson, and G. Leontidis, "Multi-source deep domain adaptation for quality control in retail food packaging," 2020, *arXiv:2001.10335*.
[331] M. Wulfmeier, A. Bewley, and I. Posner, "Incremental adversarial domain adaptation for continually changing environments," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 4489–4495.
[332] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige, S. Levine, and V. Vanhoucke, "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 4243–4250.
[333] F. Barbato, M. Toldo, U. Michieli, and P. Zanuttigh, "Latent space regularization for unsupervised domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2835–2845.
[334] D. Kothandaraman, R. Chandra, and D. Manocha, "BoMuDANet: Unsupervised adaptation for visual scene understanding in unstructured driving environments," 2020, *arXiv:2010.03523*.
[335] F. Munir, S. Azam, and M. Jeon, "SSTN: Self-supervised domain adaptation thermal object detection for autonomous driving," 2021, *arXiv:2103.03150*.
[336] M. Teichmann, M. Weber, M. Zollner, R. Cipolla, and R. Urtasun, "MultiNet: Real-time joint semantic reasoning for autonomous driving," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1013–1020.
[337] M. Sedinkina, N. Breitkopf, and H. Schütze, "Automatic domain adaptation outperforms manual domain adaptation for predicting financial outcomes," 2020, *arXiv:2006.14209*.
[338] B. Lebichot, Y.-A. L. Borgne, L. He-Guelton, F. Oblé, and G. Bontempi, "Deep-learning domain adaptation techniques for credit cards fraud detection," in *Recent Advances in Big Data and Deep Learning*. Florence, Italy: Association for Computational Linguistics, 2020, pp. 78–88.
[339] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, pp. 41–75, Jul. 1997.
[340] P. Gardner, X. Liu, and K. Worden, "On the application of domain adaptation in structural health monitoring," *Mech. Syst. Signal Process.*, vol. 138, Apr. 2020, Art. no. 106550.
[341] M. Long, J. Wang, G. Ding, S. J. Pan, and P. S. Yu, "Adaptation regularization: A general framework for transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1076–1089, May 2014.
[342] W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 766–785, Mar. 2021.
[343] J. Ren, I. Hacihaliloglu, E. A. Singer, D. J. Foran, and X. Qi, "Unsupervised domain adaptation for classification of histopathology whole-slide images," *Frontiers Bioeng. Biotechnol.*, vol. 7, pp. 1–10, May 2019.
[344] X. Tang, B. Du, J. Huang, Z. Wang, and L. Zhang, "On combining active and transfer learning for medical data classification," *IET Comput. Vis.*, vol. 13, no. 2, pp. 194–205, Mar. 2019.
[345] K. Kamnitsas, C. Baumgartner, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, A. Nori, A. Criminisi, D. Rueckert, and B. Glocker, "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, 2017, pp. 597–609.
[346] C. Liu, A. Mauricio, J. Qi, D. Peng, and K. Gryllias, "Domain adaptation digital twin for rolling element bearing prognostics," in *Proc. Annu. Conf. PHM Soc.*, 2020, pp. 1–10.
[347] A. G. Mahyari and T. Locker, "Domain adaptation for robot predictive maintenance systems," 2018, *arXiv:1809.08626*.
[348] T. Boucher, C. J. Cary, S. Mahadevan, and M. Dyar, "Aligning mixed manifolds," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 1–7.
[349] J. Yang, H. Zou, S. Cao, Z. Chen, and L. Xie, "MobileDA: Toward edge-domain adaptation," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 6909–6918, Aug. 2020.
[350] S. Mourragui, M. Loog, M. A. V. D. Wiel, M. J. T. Reinders, and L. F. A. Wessels, "PRECISE: A domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors," *Bioinformatics*, vol. 35, no. 14, pp. 510–519, 2019.
[351] X. Deng, H. L. Yang, N. Makkar, and D. Lunga, "Large scale unsupervised domain adaptation of segmentation networks with adversarial learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 4955–4958.
[352] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Nov. 2009.
[353] M. Latah and L. Toker, "Artificial intelligence enabled software-defined networking: A comprehensive overview," *IET Netw.*, vol. 8, no. 2, pp. 79–99, Mar. 2019.
[354] J. Yoo, Y. Hong, Y. Noh, and S. Yoon, "Domain adaptation using adversarial learning for autonomous navigation," 2017, *arXiv:1712.03742*.
[355] Y. Liu, L. Zhong, J. Qiu, J. Lu, and W. Wang, "Unsupervised domain adaptation for nonintrusive load monitoring via adversarial and joint adaptation network," *IEEE Trans. Ind. Informat.*, vol. 18, no. 1, pp. 266–277, Jan. 2022.
[356] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 5, pp. 1–46, Oct. 2020.
[357] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.
[358] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. A. Earnshaw, and I. S. Haque, "WILDS: A benchmark of in-the-wild distribution shifts," in *Proc. PMLR*, 2021, pp. 5637–5664.
[359] T. Talaviya, D. Shah, N. Patel, H. Yagnik, and M. Shah, "Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides," *Artif. Intell. Agricult.*, vol. 4, pp. 58–73, Jan. 2020.
[360] Forbes. (Feb. 7, 2021). *10 Ways AI Has The Potential To Improve Agriculture In 2021*. Accessed: Sep. 18, 2021. [Online]. Available: https://www.forbes.com/sites/louiscolumbus/2021/02/17/10-ways-ai-has-the-potential-to-improve-agriculture-in-2021/?sh=379355747f3b
[361] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3722–3731.
[362] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions," 2021, *arXiv:2107.13782*.
[363] Y. Zhang, R. Barzilay, and T. Jaakkola, "Aspect-augmented adversarial networks for domain adaptation," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 515–528, Dec. 2017.
[364] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofev, S. Ettinger, M. Krivoko, and A. Joshi, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2020, pp. 2446–2454.

**PEEYUSH SINGHAL** received the B.Tech. and M.Tech. degrees from the Indian Institute of Technology Bombay. He has over two decades of experience in software engineering, technology consulting, and data science and engineering. He is currently a Doctor of Philosophy (Ph.D.) Scholar at the Symbiosis Institute of Technology, Symbiosis International University, Pune, India. His research interest includes domain adaptation and its applications in various research tasks.

**RAHEE WALAMBE** (Senior Member, IEEE) received the M.Phil. and Ph.D. degrees from Lancaster University, U.K., in 2008. From 2008 to 2017, she was a Research Consultant with various organizations in the control and robotics domain. Since 2017, she has been working as an Associate Professor with the Department of Electronics and Telecommunications, Symbiosis Institute of Technology, Symbiosis International University, Pune, India. Her research interests include applied deep learning and AI in the field of robotics and healthcare.

**SHEELA RAMANNA** received the B.S. degree in electrical engineering and the M.S. degree in computer science and engineering from Osmania University, India, and the Ph.D. degree in computer science from Kansas State University, Manhattan, KS, USA. She is currently a Full Professor and the Past Chair of the Applied Computer Science Department. She is also the Co-Founder of the ACS Graduate Studies Program and the Founder of the Applied Computer Science Undergraduate Program, The University of Winnipeg. She has co-edited a book *Emerging Paradigms in Machine Learning* (Springer, 2013) with L. C. Jain and Robert Hewitt. Her research interests include fundamental and applied research in machine learning and intelligent systems. She was a recipient of multiple research grants, fellowships, and awards. She serves on the Editorial Board of *Transactions on Rough Sets* (TRS) journal and *International Journal of Rough Sets and Data Analysis*.

**KETAN KOTECHA** received the M.Tech. and Ph.D. degrees from the Indian Institute of Technology Bombay. He is currently the Head of the Symbiosis Centre for Applied AI, the Director of the Symbiosis Institute of Technology, and the Dean of the Faculty of Engineering, Symbiosis International University, Pune, India. He is also an Expert in artificial intelligence and deep learning. He has published widely in a number of excellent peer-reviewed journals on various topics ranging from cutting-edge AI, education policies, teaching-learning practices, and AI for all. He was a recipient of multiple international research grants and awards.